[illegible]

Technology 2000

(NASA-CP-3109-Vol-2) TECHNOLOGY 2000,
VOLUME 2 (NASA) 369 p CSCL 05A
N91-24041
--THRU--
N91-24083
Unclas
H1/99 0003635

Proceedings of a conference held at
the Washington Hilton Hotel and Towers
Washington, D.C.
November 27-28, 1990

NASA

NASA Conference Publication 3109, Vol. 2

Technology 2000

Proceedings of a conference held at
the Washington Hilton Hotel and Towers
Washington, D.C.
November 27–28, 1990

NASA
National Aeronautics and
Space Administration
Office of Management
Scientific and Technical
Information Division
1991

TECHNOLOGY 2000

SYMPOSIA
PROCEEDINGS

Presented November 27 & 28, 1990
Washington, D.C.

TECHNOLOGY 2000 was the first major industrial conference and exposition spotlighting NASA technology and technology transfer. It's purpose was, and continues to be, to increase awareness of existing NASA-developed technologies that are available for immediate use in the development of new products and processes, and to lay the groundwork for the effective utilization of emerging technologies.

In addition to an exhibit showcasing the products and technologies available for sale or license, TECHNOLOGY 2000 featured 16 concurrent technical sessions in which 97 papers were presented. The program included symposia on Computer Technology and Software Engineering; Human Factors Engineering and Life Sciences; Information and Data Management; Materials Science; Manufacturing and Fabrication Technology; Power, Energy and Control Systems; Robotics; Sensors and Measurement Technology; Artificial Intelligence; Environmental Technology; Optics and Communications; and, Superconductivity.

We are pleased to provide the proceedings from these sessions. They have been published in two volumes. Volume One contains the papers presented Tuesday, November 27th; Volume Two contains the papers presented Wednesday, November 28th.

This is Volume Two. It consists of 46 papers. The papers appear in the order in which they were presented at TECHNOLOGY 2000.

For information regarding additional copies, please contact:

ASSOCIATED BUSINESS PUBLICATIONS
41 EAST 42ND STREET
NEW YORK, NY 10017-5391
(212) 490-3999

PRECEDING PAGE BLANK NOT FILMED

TABLE OF CONTENTS

WEDNESDAY NOVEMBER 28TH

SESSION A: ARTIFICIAL INTELLIGENCE

Intelligent Computer-Aided Training And Tutoring. 3
CLIPS: A Tool For The Development And Delivery Of Expert Systems. 11
Distributed, Cooperating Knowledge-Based Systems. 16
Electronic Neuroprocessors. 26
Intelligent Vision System For Autonomous Vehicle Operations. 34

SESSION B: COMPUTER TECHNOLOGY AND SOFTWARE ENGINEERING (PART 2)

Software Reengineering. 47
Transportable Applications Environment (TAE) Plus: A NASA User Interface Development And Management System. 56
Applications Of Fuzzy Logic To Control And Decision Making. 67
Genetic Algorithms. 76
Integrated Vertical Bloch Line (VBL) Memory. 86

SESSION C: ENVIRONMENTAL TECHNOLOGY

Physical/Closed-Loop Water Recycling. 95
Water Quality Analyzer. 106
New Research On Bioregenerative Air/Water Purification Systems. 107
Environmental And Facilities Management System (EMFS). 114
The Land Analysis System. 115
Potential Commercial Use Of EOS Remote Sensing Products. 122

SESSION D: HUMAN FACTORS ENGINEERING AND LIFE SCIENCES (PART 2)

Simulation Of Blood Flow Through An Artificial Heart. 133
Three-Dimensional Structure Of Human Serum Albumin. 146
A Noninvasive Measure Of Minerals And Electrolytes In Tissue. 151
Oxygen Production Using Solid-State Zirconia Electrolyte Technology. 155
Monitoring And Control Technologies For Bioregenerative Life Support Systems. 161

SESSION E: MATERIALS SCIENCE (PART 2)

Silicon Carbide, An Emerging High-Temperature Semiconductor. 171
Flexible Fluoropolymer-Filled Protective Coatings. 179
A Conformal Oxidation-Resistant, Plasma-Polymerized Coating. 185
Flame-Retardant Composite Materials. 192
Superplastic Forming Of Al-Li Alloys For Lightweight, Low-Cost Structures. 200
Localized Corrosion Of High-Performance Metal Alloys In An Acid/Salt Environment. 210

PRECEDING PAGE BLANK NOT FILMED

TABLE OF CONTENTS

WEDNESDAY NOVEMBER 28TH (cont'd)

SESSION F: OPTICS AND COMMUNICATIONS

Digital Codec For Real-Time Processing Of Broadcast-Quality Video Signals. 221
Recent Advances In Coding Theory For Near-Error-Free Communications. 229
Microwave Integrated Circuits For Space Applications. 242
Optical Communications For Space Missions. 249
Optical Shutter Switching Matrix. 259
Fiber Optical Tactical Local Area Network (FOTLAN). 268
High-Precision Applications Of The Global Positioning System. 276

SESSION G: SENSORS AND MEASUREMENT TECHNOLOGY (PART 2)

Quantitative Nondestructive Evaluation - Requirements For Tomorrow's Reliability. 287
Robotic Control And Inspection Verification. 298
The Transfer Of A Technology To Measure Skin Burn Depth In Humans. 308
Frequency Domain Laser Velocimeter Signal Processor. 315
A Field-Deployable Digital Acoustic Measurement System. 325
Laser-Optical Disk Position Encoder With Active Heads. 333

SESSION H: SUPERCONDUCTIVITY

Applications Of High-Temperature Superconductors. 343
Superconducting Microwave Electronics At NASA Lewis. 346
Superconductive Wires And Devices For Cryogenic Applications. 355
Melt-Processed Bulk Superconductors Fabrication And Characterization For Power And Space Applications. 356
Large Gap Magnetic Suspension System Ground-Based Experiment. 365

SESSION A - ARTIFICIAL INTELLIGENCE

Wednesday November 28, 1990

- **Intelligent Computer-Aided Training And Tutoring**
- **CLIPS: A Tool For The Development And Delivery Of Expert Systems**
- **Distributed, Cooperating Knowledge-Based Systems**
- **Electronic Neuroprocessors**
- **Intelligent Vision System For Autonomous Vehicle Operations**

N91

24042

UNCLAS

N91-24042

Intelligent Computer-Aided Training and Tutoring

R. Bowen Loftin, Ph.D.
University of Houston-Downtown
One Main Street
Houston, Texas

Robert T. Savely, Chief
PT4/Software Technology Branch
NASA/Johnson Space Center
Houston, Texas

INTRODUCTION

Motivation

Training is a major endeavor in all modern societies: new personnel must be trained to perform the task(s) which they were hired to perform, continuing personnel must be trained to upgrade or update their ability to perform assigned tasks, and continuing personnel must be trained to tackle new tasks. Common methods include training manuals, formal classes, procedural computer programs, simulations, and on-the-job training. The latter method is particularly effective in complex tasks where a great deal of independence is granted to the task performer. Of course, this training method is also the most expensive and may be impractical when there are many trainees and few experienced personnel to conduct on-the-job training.

NASA's training approach has focussed primarily on on-the-job training in a simulation environment for both crew and ground-based personnel. This process worked relatively well for both the Apollo and Space Shuttle programs. Space Station Freedom and other long range space exploration programs coupled with limited resources dictate that NASA explore new approaches to training for the 1990s and beyond.

This report describes specific autonomous training systems based on artificial intelligence technology for use by NASA astronauts, flight controllers, and ground-based support personnel that demonstrate an an alternative to current training systems. In addition to these specific systems, the evolution of a general architecture for autonomous intelligent training systems that integrates many of the features of "traditional" training programs with artificial intelligence techniques is presented. These Intelligent Computer-Aided Training (ICAT) systems would provide, for the trainee, much of the same experience that could be gained from the best on-the-job training. By integrating domain expertise with a knowledge of appropriate training methods, an ICAT session should duplicate, as closely as possible, the trainee undergoing on-the-job training in the task environment, benefiting from the full attention of a task expert who is also an expert trainer. Thus, the philosophy of the ICAT system is to emulate the behavior of an experienced individual devoting his full time and attention to the training of a novice—proposing challenging training scenarios, monitoring and evaluating the actions of the trainee, providing meaningful comments in response to trainee errors, responding to trainee requests for information, giving hints (if appropriate), and remembering the strengths and weaknesses displayed by the trainee so that appropriate future exercises can be designed.

Background

Since the 1970s a number of academic and industrial researchers have explored the application of artificial intelligence concepts to the task of teaching a variety of subjects [1, 2, 3] (*e.g.*, computer programming in Lisp [4, 5] and Pascal [6], economics [7], and, geography [8]). The earliest published reports which suggested the applications of artificial intelligence concepts to teaching tasks appeared in the early 1970s [8, 9]. Hartley and Sleeman [9] actually proposed an architecture for an intelligent tutoring system. However, it is interesting to note that, in the sixteen years which have passed since the appearance of the Hartley and Sleeman proposal, no agreement has been reached among researchers on a general architecture for intelligent tutoring systems [2].

Along with the extensive work on intelligent tutoring systems for academic settings has come the development of systems directed at training. Among these are Recovery Boiler Tutor [10], SOPHIE [11], and STEAMER [12]. These differ from the tutoring systems mentioned above in providing a simulation model with which the student or trainee interacts. Although these intelligent training systems each use the

PRECEDING PAGE BLANK NOT FILMED

interactive simulation approach, they each have very different internal architectures. Further, there appears to be no agreement, at present, on a general architecture for such simulation training systems. The work reported here builds on these previous efforts and our own work [13, 14, 15, 16] to develop specific intelligent training systems as well as a general approach to the design of intelligent training systems which will permit the production of such systems for a variety of tasks and task environments with significantly less effort that is now required to "craft" such a system for each application.

APPLICATIONS

The ICAT architecture was originally applied to a training system for NASA flight controllers learning to deploy satellites from the Space Shuttle. The same architecture has been used in the construction of ICAT systems for training astronauts for Spacelab missions and engineers who test the Space Shuttle main propulsion system. Although these tasks are quite different and are performed in very dissimilar environments, the same system architecture has proven to be adaptable to each. Below is a brief summary of the specific systems that have been built or are currently under development:

PD/ICAT: [Payload-assist module Deploys/ICAT System]

A comprehensive intelligent computer-aided training system for use by Flight Dynamics Officers in learning to deploy PAM (Payload-Assist Module) satellites from the Space Shuttle. PD/ICAT contains four expert systems that cooperate via a blackboard architecture.

VVL/ICAT: [Vacuum Vent Line/ICAT System]

A PC-based intelligent computer-aided training system for use by mission and payload specialists in learning to perform fault detection, isolation, and reconfiguration on the Spacelab VVL system. VVL/ICAT consists of an integrated expert system and graphical user interface.

MPP/ICAT: [Main Propulsion Pneumatics/ICAT System]

A comprehensive intelligent computer-aided training system for use by test engineers at NASA/Kennedy Space Center in learning to perform testing of the Space Shuttle Main Propulsion Pneumatics system. MPP/ICAT is currently under development and makes use of the same architecture as PD/ICAT.

IPS/ICAT: [Instrument Pointing System/ICAT System]

A comprehensive intelligent computer-aided training system for use by payload and mission specialists at NASA/Johnson Space Center and Marshall Space Flight Center in learning to utilize the IPS on Spacelab missions. IPS/ICAT is currently under development and makes use of the same architecture as PD/ICAT.

A GENERAL ARCHITECTURE FOR INTELLIGENT TRAINING SYSTEMS

The projects described in the previous section have served as vehicles to aid in the design and refinement of an architecture for intelligent training systems that has significant domain-independent elements and is generally applicable to training in procedural tasks common to the NASA environment. The ICAT system architecture is modular and consists of five basic components:

- A user interface that permits the trainee to access the same information available to him in the the task environment and serves as a means for the trainee to take actions and communicate with the intelligent training system.

- A domain expert which can carry out the task using the same information that is available to the trainee and which also contains a list of "mal-rules" (explicitly identified errors that novice trainees commonly make).

- A training session manager which examines the actions taken by the domain expert (of both correct and incorrect actions in a particular context) and by the trainee and takes appropriate action(s). [17]

- A trainee model which contains a history of the individual trainee's interactions with the system together with summary evaluative data.

- A training scenario generator that designs increasingly-complex training exercises based on the knowledge of the domain expert, the current skill level contained in the trainee's model, and any weaknesses or deficiencies that the trainee has exhibited in previous interactions. [18, 19]

Figure 1 contains a schematic diagram of the ICAT system. Note that provision is made for the user to interact with the system in two distinct ways and that a supervisor may also query the system for evaluative data on each trainee. The blackboard serves as a common repository of facts for all five system components. With the exception of the trainee model, each component makes assertions to the blackboard, and the expert system components look to the blackboard for facts against which their rules pattern match. A comprehensive effort has been made to clearly segregate domain-dependent from domain-independent components.

The ICAT architecture described above was originally implemented in a Symbolics 3600 Lisp environment using Inference Corporation's ART for the rule-based components. The architecture is currently available for unix workstations. The user interface is implemented in X-Windows, the rule-based components in CLIPS [CLIPS is the acronym for a NASA-developed expert system shell written in C], and supporting code in C.

TRAINING PERFORMANCE

The original system developed with this architecture (PD/ICAT) has been used by both expert and novice flight controllers at NASA/Johnson Space Center. An extensive investigation of the performance of novices using the system has been conducted. Figure 2 shows two measures of performance: (1) the time required to perform the nominal task as a function of the number of training experiences and (2) the number of errors made during the performance of the nominal task as a function of the number of training experiences. It is interesting to note that, although the novices used in this investigation had very different levels of prior experience related to the task, all novices rapidly approached the same level of proficiency.

A TECHNOLOGY SPINOFF: THE INTELLIGENT PHYSICS TUTOR

Introduction

The integration of the computer into the K-12 instructional program began in the 1960s and has accelerated with the availability of inexpensive computing hardware and a growing amount of useful instructional software [20]. The bulk of the computer-aided instruction (CAI) available today is limited to rather simple programs that are useful for drill-and-practice, automated "page-turning", and the administration of objective examinations. Only a small percentage of the educational software available today for grades K-12 uses simulation, extensive branching to diagnose and remediate, and/or artificial intelligence (AI) technology [21].

Objectives

The ICAT technology described above has been brought to bear on a specific tutoring task of interest to the nation's educational institutions: the development of an intelligent tutoring system (ITS) for use in a high school or introductory college physics course. The goal of this ITS is not the conveyance of facts and concepts but rather the transfer of problem solving skills to the student. Ultimately, this project will not only produce a useful teaching aid for students enrolled in high school or introductory college physics courses, but will also provide a development structure suitable for building additional intelligent tutors for other academic subjects which require the application of problem solving skills (*e.g.*, mathematics, chemistry, and engineering).

Technology Transfer

The work described above led the NASA/JSC Office of Technology Utilization to suggest, in February, 1988, that this technology be applied to the development of an intelligent tutoring system suitable for use in the nation's educational institutions. The outcome of this technology transfer project would be a highly interactive and "intelligent" program for tutoring high school and college students in solving physics problems. This product could be mass produced and delivered economically to high schools and colleges nationwide. In addition, the methodologies employed and much of the software developed could be used to produce intelligent tutors for other problem solving domains. The effect of this technology transfer would clearly be profound and is entirely consistent with NASA's charter.

Project Description

Background. By the beginning of this decade a number of researchers had attempted to develop intelligent tutors for selected tutoring or training tasks. Specific systems such as GUIDON (medical diagnosis) [22, 23], SOPHIE (electronic troubleshooting) [11], STEAMER (naval steam propulsion systems) [12, 24], WUSOR (reasoning) [25], PROUST (Pascal programming) [6, 26], the GEOMETRY Tutor (plane geometry) [27, 28], and the LISP Tutor (LISP Programming) [4] have been built and tested.

Perhaps the most successful of these, the LISP Tutor, has been shown to have a significant effect on student mastery of a skill. For example, in one controlled experiment, two groups of students attended the same lectures on LISP programming, and one group completed the exercises in the traditional manner while the other group used the LISP Tutor. The students using the LISP Tutor spent 30% less time on the exercises and scored 43% better on a post-test than those not using the LISP Tutor [5]. Further classroom use of the LISP Tutor at Carnegie-Mellon University for a one-semester course in LISP showed that students were able, as a whole, to achieve a full letter grade improvement in their final course grade by using the LISP Tutor when compared with previous classes that did not use the tutor [29]. Interestingly, poor students demonstrated the most significant performance improvement [5].

Approach. An ITS containing at least one semester of exercises for the first course in high school and introductory college physics is now under development. If time is available, additional exercises appropriate for the second semester will also be developed. The ITS will be designed for use by students concurrently enrolled in the standard high school or introductory college physics course and will be used to tutor them in physics problem solving. In order to provide an interactive environment suitable for detailed simulations of physical phenomena, the ITS will be delivered as a part of a workstation built around the Apple Macintosh II computer. The interface will make full use of the Macintosh II's capabilities—using high-resolution graphics, color and sound to deliver a sophisticated real-time simulation as a means of enhancing the students' ability to relate the tutorial environment to that of the laboratory and the "real world". The strategy (adapted, in part, from the LISP Tutor) of examining user input continuously and providing immediate feedback upon detecting errors is utilized. The intelligent physics tutor is not intended to replace a human instructor or to replace the existing program of physics instruction. Instead, the tutor that is under development will provide an interactive environment for the application of physics concepts to the solution of problems. The principal goal of the tutor will be to enable the average student to efficiently acquire problem solving skills necessary for successful mastery of high school or introductory college physics. It is anticipated that a large fraction of such physics classes could be served by the tutor, freeing the human instructor to work more closely with the slower students as well as with the more advanced students. The tutorial lessons will be integrated with the lecture/laboratory portions of the typical course, and students will have independent access to the tutor for completion of their homework.

Since the intent of the tutor is to complement an existing textbook and course in which concepts will be logically introduced, it is essential that the tutor be adaptable to a variety of texts. To this end the tutor lessons (here "lesson" refers to a sequence of student-tutor interactions that lead to student mastery of problem solving skills that pertain to a specific concept) are keyed to physics concepts. The instructor will be provided the capability of assembling the lessons into an order consistent with the curriculum followed in the classroom and of establishing appropriate dependencies among the lessons. In addition, the instructor will have the means of choosing terms and symbols compatible with their textbook and/or preferences. Each instance of the physics tutor will contain a "global" strategy to govern the student's progression from lesson to lesson, but the tutor strategy and the semantics used in a given lesson may be altered by the instructor. The lessons for these topics will vary in length according to the complexity of problems associated with each topic. For example, the lesson for uniform acceleration in one-dimension will contain many more exercises, of varying difficulty, for the student than the lesson related to kinetic energy for rotational motion.

For each tutor lesson the student progresses from the solution of a "one-step" problem in which the available facts and the required item(s) are clearly delineated and a single relation allows the student to obtain the required result, to multi-step problems in which intermediate results or interdependencies are required to obtain the information necessary for the final solution, to problems in which the student must, given the required item(s), obtain the necessary data through the observation of a simulation.

In the "text-only" problems the student, in initial tutoring sessions, may be provided with a structured environment within which to solve the problem. The listing of given information and the identification of the required item(s) can be enforced by the tutor. The student is provided regions in which to assemble lists of given and required data obtained from the problem or implied by the problem. By using a natural and powerful interface, a student can use assemble his or her solution on the computer screen as if working the problem on a piece of notebook paper. At any time during the interaction the tutor can intervene by "popping-up" a window and/or providing audible feedback. In addition, the student can obtain assistance (examples, hints, etc.) by using the help menu item.

For the most advanced exercises in many lessons the tutor can lead the student in abstracting real-world problems in a manner that facilitates their solution via the methodologies conveyed in the preceding exercises. Experience has shown that most students have great difficulty in reading text describing a problem (or even viewing a photograph, detailed drawing, or video image) and producing an abstract representation of the problem that fits the problem solving patterns they have been shown by an instructor or in a textbook. Consider, for example, a video segment of a skier on a uniform slope. Most novice problem solvers have difficulty in abstracting such a scene by drawing a free-body diagram showing the skier as a point mass with the three forces (gravity, friction, and the normal force) acting on him or her. Such a step, however, is essential in order to apply Newton's Second Law and and determine the skier's acceleration. The tutorial strategy in this case is accomplished by providing video stills or video segments of a physical situation and overlaying a graphical abstraction. By causing the video image to fade and the graphical abstraction to remain, students can rapidly gain experience in abstracting the essential elements of a physical description. Student mastery can be tested by allowing the student to construct the appropriate graphical abstraction from a video image and, eventually, from a written problem statement.

Error detection and remediation occur at a local level with each student action compared to expected correct and incorrect actions. Based on the nature of each student error, appropriate feedback is given to enable the student to understand and correct his or her error. A global tutor strategy draws on a student model which identifies lacks of skills and knowledge demonstrated by the student in attempting to solve problems within the tutor environment. The number, type, and order of the exercises encountered by the student is determined by heuristic rules which examine the student model and the performance of the student on the exercises completed in the current lesson. The student model is also used to determine the way in which assistance is provided to the student. That is, the type, length, and tenor of messages can be tailored to the student based on the history of their experience with the tutor as contained in the student model.

As the tutor's development has proceeded, a number of components have been created which can serve as the basis for the production of tutors for problem-solving domains other than physics. The interface development has produced a set of basic objects that can also be used to build interfaces for tutors in domains other than physics. An environment for encoding the domain knowledge and the global tutoring strategy has been developed to facilitate the creation of the rule-based portions of the tutor. This environment will also be adaptable to the creation of rule-based components for other tutoring systems.

The first year of this project was dedicated to the development of prototypes in cooperation with a physics teacher from a local school district. The prototypes have been in use in a local high school since April, 1989, and a evaluation/ development/refinement feedback loop has been established. Beginning in September, 1990, a semester of physics exercises is being used to begin formal testing with a students in League City, Texas and Columbus, Ohio. Finally, during the third year of the project, refinement of the ITS and its further extension will occur.

The ITS design and evaluation has been carried out by Dr. R. Bowen Loftin (University of Houston-Downtown), Dr. Steve Brown (Cognitive Systems Technologies), and Ms. Beverly Lee (a local high school physics teacher). Computer Sciences Corporation has provided personnel for the principal code development activity. Gary Riley and Brian Donnell of the Software Technology Branch at NASA/Johnson Space Center have also contributed to the design and coding of some tutor elements.

CONCLUSIONS

A general architecture for ICAT systems has been developed and applied to the construction of three ICAT systems for very different tasks. Use by novices of an ICAT application built upon this architecture has shown impressive trainee performance improvements. With further refinement and extension, this architecture promises to provide a common foundation upon which to build intelligent training systems for many tasks of interest to the government, military, and industry. The availability of a robust architecture that contains many domain-independent components serves to greatly reduce the time and cost of developing new ICAT applications. As an added benefit to the nation, a technology spinoff project has emerged from this activity and promises to make a significant contribution to the secondary and post-secondary education.

REFERENCES

[1] Sleeman, D. and Brown, J.S., eds., *Intelligent Tutoring Systems*, Academic Press, London, 1982.

[2] Yazdani, M. "Intelligent Tutoring Systems Survey," *Artificial Intelligence Review*, Vol. 1, p. 43, 1986.

[3] Wenger, E., *Artificial Intelligence and Tutoring Systems*, Morgan Kaufmann Publishers, Los Altos, CA, 1987.

[4] Anderson, J. R. and Reiser, B. J., "The LISP Tutor," *BYTE*, 159 (April, 1985).

[5] Anderson, J. R., Boyle, C. F., and Reiser, B. J., "Intelligent Tutoring Systems," *Science 228*, 456 (1985).

[6] Johnson, W. L. and Soloway, E., "PROUST," *Byte 10* (4), 179 (April, 1985).

[7] Shute, V. and Bonar, J.G., An Intelligent Tutoring System for Scientific Inquiry Skills, *Proceedings of the Eighth Cognitive Science Society Conference* Amherst, MA, 1986, pp. 353-370.

[8] Carbonell, J.R., "AI in CAI: An Artificial Intelligence Approach to CAI," *IEEE Transactions on Man-Machine Systems 11*(4), 190 (1970).

[9] Hartley, J. R. and Sleeman, D. H., "Towards Intelligent Teaching Systems," *International Journal of Man-Machine Studies 5*, 215 (1973).

[10] Woolf B.P., Blegen, D., Jansen, J.H., and Verloop, A., Teaching a Complex Industrial Process, *Proceedings of the National Conference on Artificial Intelligence*, Vol. II, Austin, TX, 1986, pp. 722-728.

[11] Brown, J. S., Burton, R. R., and de Kleer, J., "Pedagogical, Natural Language and Knowledge Engineering Techniques in SOPHIE I, II, and III," in D. Sleeman and J. S. Brown (eds.), *Intelligent Tutoring Systems* (London: Academic Press, 1982), p. 227.

[12] Hollan, H. D., Hutchins, E. L., and Weitzman, L., "Steamer: An Interactive Inspectable Simulation-based Training System," *AI Magazine*, Summer, 1984, pp. 15-27.

[13] Loftin, R.B., Wang, L., Baffes, P., and Rua, M., An Intelligent Computer-Aided Training System for Payload-Assist Module Deploys, *Proceedings of the First Annual Workshop on Space Operations*,

Automation & Robotics (SOAR '87), NASA/Johnson Space Center, Houston, TX, August 4-6, 1987.

[14] Loftin, R.B., Wang, L., Baffes, P., and Hua, G., An Intelligent Training System for Space Shuttle Flight Controllers, *Telematics and Informatics 5* (3), 151-161 (1988).

[15] Loftin, R.B., Wang, L., Baffes, P., and Hua, G., "An Intelligent System for Training Space Shuttle Flight Controllers in Satellite Deployment Procedures," *Machine Mediated Learning 3*(1), 41 (1989).

[16] Loftin, R.B., Wang, L., Baffes, P., and Hua, G., "An Intelligent Training System for Space Shuttle Flight Controllers" in *Innovative Applications of Artificial Intelligence*, edited by H. Schorr and A. Rappaport (Menlo Park, CA: AAAI Press, 1989), pp. 15-24.

[17] Loftin, R.B., Baffes, P., and Wang, L, An Approach to the Diagnosis of Trainee Errors in an Intelligent Training System, unpublished report.

[18] Loftin, R.B., Wang, L., and Baffes, P., "Simulation Scenario Generation for Intelligent Training Systems," *Proceedings of the Third Artificial Intelligence and Simulation Workshop*, August 22, 1988, St. Paul, MN, pp. 69-74.

[19] Loftin, R.B., Wang, L., and Baffes,P., "Intelligent Scenario Generation for Simulation-Based Training," *Proceedings of the American Institute for Aeronautics and Astronautics Computers in Aerospace VII Conference*, Monterey, CA, October 3-5, 1989, pp. 581-588.

[20] Bork, A., *Learning with Computers*, Digital Press, Bedford, MA, 1981.

[21] Bork, A., "Nontrivial, Nonintelligent Computer-Based Learning," Technical Report, Educational Technology Center, University of California, Irvine, CA, 1986.

[22] Clancey, W. J., "GUIDON," *J. Computer-Based Instruction 10*(1), 8 (1983).

[23] Clancey, W. J., Shortliffe, E. H., and Buchanan, B. G., "Intelligent Computer-Aided Instruction for Medical Diagnosis," in W. J. Clancey and E. H. Shortliffe (eds.), *Readings in Medical Artificial Intelligence: The First Decade* (Reading, MA: Addison-Wesley, 1984).

[24] Williams, M. D., Hollan, J. D., and Stevens, A. L., "An Overview of STEAMER: an Advanced Computer-assisted Instruction System for Propulsion Engineering," *Behavior Research Methods and Instrumentation 13*(2), 85 (1981).

[25] Carr, B., "Wusor-II: a Computer-aided Instruction Program with Student Modeling Capabilities," AI Lab Memo 417, Massachusetts Institute of Technology, Cambridge, MA, 1977.

[26] Johnson, W. L. and Soloway, E., "PROUST: Knowledge-based Program Debugging," *Proceedings of the Seventh International Software Engineering Conference* , Orlando, Florida, 1984, pp. 369-380.

[27] Farrell, R. G., Anderson, J. R., and Reiser, B., "An Interactive Computer-Based Tutor for LISP," *Proceedings of the American Association for Artificial Intelligence* Conference (Los Altos, CA: Morgan-Kaufman, 1984).

[28] Anderson, J. R., Boyle, C. F., and Yost, G., "The Geometry Tutor," *Proceedings of the Ninth International Joint Conference on Artificial Intelligence* (Los Altos, CA: Morgan-Kaufman, 1985), pp. 1-7.

[29] Anderson, J. R. and Skwarecki, E., "The Automated Tutoring of Introductory Programming," *Communications of the ACM 29*(9), 842 (1986).

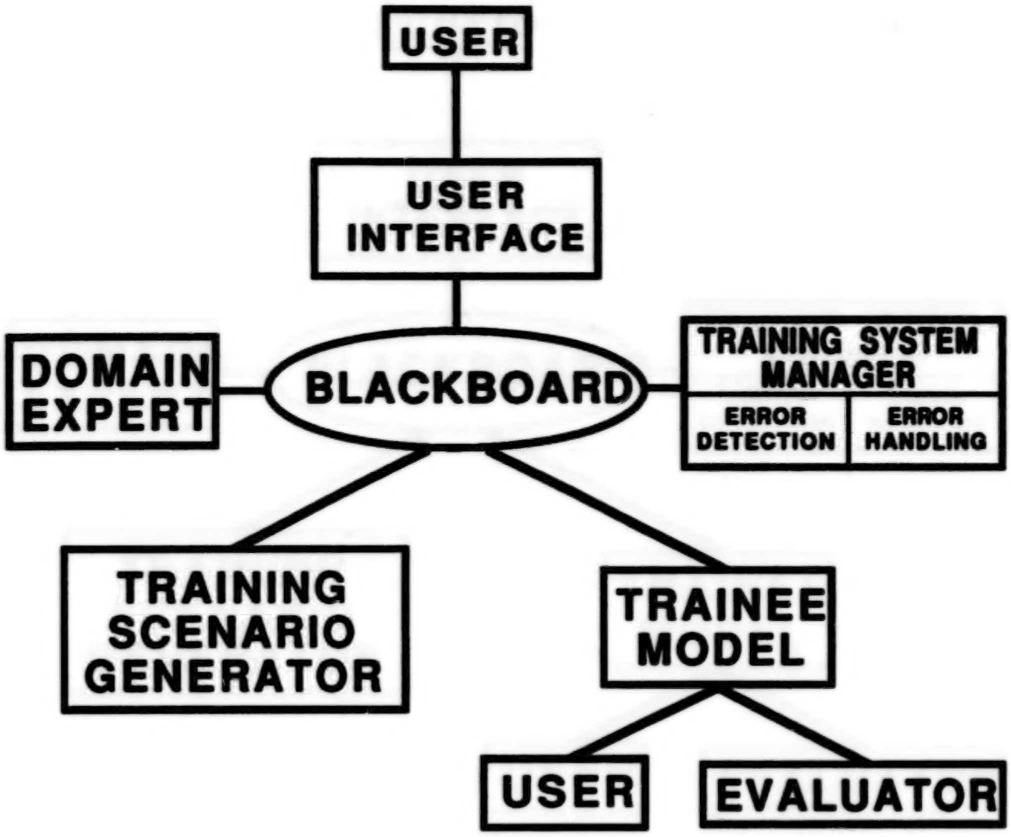

USER
USER INTERFACE
DOMAIN EXPERT
BLACKBOARD
TRAINING SYSTEM MANAGER
ERROR DETECTION
ERROR HANDLING
TRAINING SCENARIO GENERATOR
TRAINEE MODEL
USER
EVALUATOR

Figure 1. A Schematic Diagram of the General ICAT Architecture

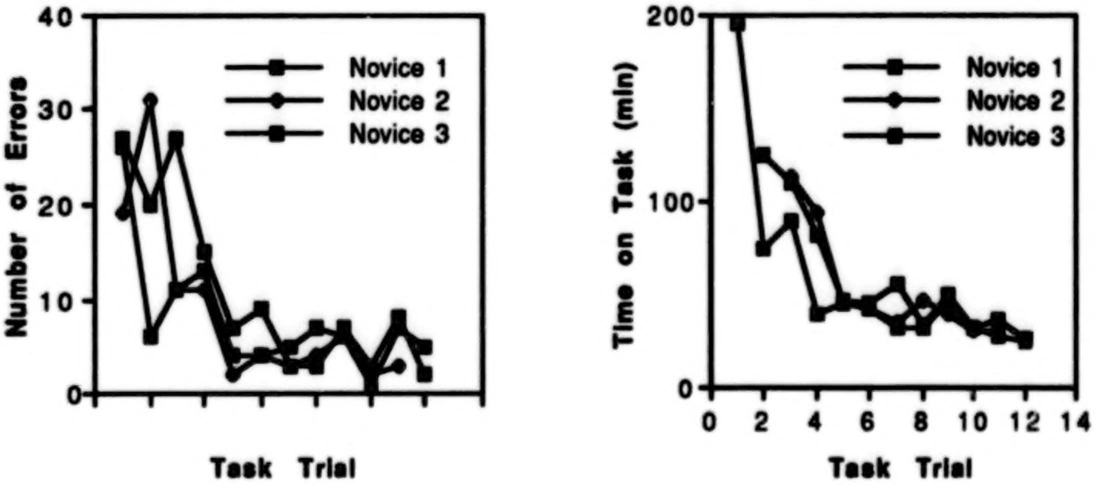


Figure 2. Performance of Novices Using the PD/ICAT System

N91
24043
UNCLAS

N91-24043

CLIPS: A Tool for the Development and Delivery of Expert Systems

Gary Riley
Software Technology Branch
NASA Johnson Space Center
Mail Stop PT4
Houston, TX 77058

ABSTRACT

The 'C' Language Integrated Production System (CLIPS) is a forward chaining rule-based language developed by the Software Technology Branch at the Johnson Space Center. CLIPS provides a complete environment for the construction of rule-based expert systems. CLIPS was designed specifically to provide high portability, low cost, and easy integration with external systems. Other key features of CLIPS include a powerful rule syntax, an interactive development environment, high performance, extensibility, a verification/validation tool, extensive documentation, and source code availability. The current release of CLIPS, version 4.3, is being used by over 2,500 users throughout the public and private community including: all NASA sites and branches of the military, numerous federal bureaus, government contractors, 140 universities, and many companies.

INTRODUCTION

Expert system technology is a major subset of Artificial Intelligence and has been aggressively pursued by researchers since the early 1970's. In the last few years, both government and commercial application developers have given expert systems considerable attention as well. An entire industry has grown to support the development of expert system tools and applications, with a wide variety of both hardware and software products now available. The availability of expert system tools has greatly reduced the effort and cost involved in developing an expert system.

Despite all this, expert systems have generally failed to make a major impact in application environments. This failure has stemmed from tool vendor's overemphasis on expert system development environments to the detriment of options for delivery of expert systems and training in expert system technology. Viable delivery options are necessary to field expert systems. Training options in expert system technology are necessary for the widest possible dissemination of this technology.

The 'C' Language Integrated Production System (CLIPS) is a forward chaining rule-based production system developed by the Software Technology Branch at NASA/Johnson Space Center. Version 1.0 of CLIPS, developed in the spring of 1985 in a little over two months, accomplished two major goals. The first of these goals was to gain useful insight and knowledge about the construction of expert system tools and to lay the groundwork for future versions. The second of these goals was to address the delivery problems of integrating and embedding expert systems into conventional environments. Version 1.0 successfully demonstrated the feasibility of continuing the project.

Subsequent development of CLIPS greatly improved its portability, performance, and functionality. A reference manual [1], architecture manual [2], and user's guide [3] were written. The first release of CLIPS, version 3.0, was in July of 1986. The latest version of CLIPS, version 4.3, was completed in June of 1989. A version of CLIPS developed entirely in Ada and fully syntax compatible with the C version of CLIPS has also been developed. CLIPS is currently available through COSMIC (see appendix).

DELIVERY

CLIPS was designed, in part, to solve problems related to the delivery of expert systems. CLIPS addresses several issues key to the delivery problem. Among these issues are: the ability to run on conventional hardware; the ability to run on a wide variety of hardware platforms; the ability to be integrated with and embedded within conventional software; low-cost delivery options; the ability to separate the development environment from the delivery environment (i.e. run-time modules); the ability to run efficiently (both speed and memory), and migration paths from development to delivery environments.

One major requirement for a delivery tool is the ability to run on conventional hardware. Portability of the expert system tool code insures the ability to deliver on a wide range of hardware from microcomputers to minicomputers to mainframes. Because CLIPS is written in C and special care was taken to preserve portability, CLIPS is able to provide expert system technology on a wide variety of conventional computers. CLIPS has been hosted on over a dozen brands of computer systems ranging from microcomputers to mainframes without code changes. To maintain portability, CLIPS utilizes the concept of a portable kernel. The kernel represents a section of code which utilizes no machine dependent features. To provide machine dependent features, such as windowed interfaces or graphics editors, CLIPS provides fully documented software hooks which allow machine dependent features to be integrated with the kernel.

At the time of its development, CLIPS was one of the few tools that was written in C and capable of running on a wide variety of conventional platforms. Most state-of-the-art expert system software tools at that time were based in LISP and ran only on specialized LISP hardware, such as the Symbolics or TI Explorer. In recent years, many tool vendors have migrated their products to workstations and PCs and numerous products are now available in C, Ada, and other conventional languages.

The ability to integrate with and embed within existing code is an important feature for a delivery tool. Integration guarantees that an expert system does not have to be relegated to performing tasks better left to conventional procedural languages. It also allows existing conventional code to be utilized. The capability to be embedded allows an expert system to be called as a subroutine (representing perhaps only one small part of a much larger program). Many tools view themselves as the "master" program and only permit control to be passed to other programs through them. CLIPS allows integration with C programs as well as integration with other languages such as FORTRAN and ADA. In addition, many functions are provided which allow CLIPS to be manipulated externally. Because the source code is available, CLIPS can be modified or tailored to meet a specific user's needs.

Applications should be delivered as economically as possible. Many tools require the entire development environment to run an application. This necessitates buying a new copy of the tool for every delivered application. Some tools provide the capability to generate run-time modules. These run-time modules are basically equivalent to the executable modules generated by compilers for procedural languages. Run-time modules allow the unneeded functionality and information associated with the development environment to be stripped away from the delivery environment. This is a desirable characteristic, but for many tools, each copy of a run-time module must be purchased.

CLIPS effectively addresses the problems of low cost delivery. The cost for CLIPS source code is $250. This initial cost provides unlimited copies of CLIPS for delivery, development, and training. In addition, CLIPS also provides the capability to generate run-time modules.

Another key feature for a delivery tool is efficiency. CLIPS is based on the Rete algorithm [4] which is an extremely efficient algorithm for pattern matching. CLIPS version 4.3 compares quite favorably to other commercially available expert system tools based on the Rete algorithm.

CURRENT USES

Although CLIPS was originally developed to help in the construction of aerospace related expert systems, it has been put to widespread usage in a number of fields. The current release of CLIPS, version 4.3, is being used by over 2,500 users throughout the public and private community including: all NASA sites and branches of the military, numerous federal bureaus, government contractors, 140 universities, and many companies. At the First CLIPS Conference held in August 1990, over 80 papers were presented on a diverse range of topics. In addition to aerospace and engineering applications, some other examples of CLIPS applications include: software engineering [5, 6], networking [7, 8], medical and biological [9, 10], and agricultural [11, 12].

EXTENSIONS

One of the key appeals of the CLIPS language results from the availability of the approximately 25,000 lines of CLIPS source code. Because the development of an expert system tool can require many man-years, the benefits of using CLIPS as a starting point for research and the creation of special purpose expert system tools cannot be understated. CLIPS users have enjoyed a great deal of success in adding their own extensions to CLIPS due to the source code availability and its open architecture. Some of the many extensions added by CLIPS users include: an SQL interface [13, 14], neural network extensions [15, 16], blackboard extensions [17], parallel and distributed processing [18, 19, 20, 21, 22], and object-oriented programming extensions [23, 24, 25].

Because CLIPS was written to be portable, its standard interface is a simple, text-oriented, command prompt. However, CLIPS provides the capability to layer an interface on top of CLIPS to provide the advantages of a windowed, mouse-driven, menu-oriented user interface. Currently, two such interfaces are provided with CLIPS: one for the Apple Macintosh family of computers and one for IBM PC compatible computers. An X-Windows interface is currently under development. Many users have also developed their own interfaces and interface extensions [26, 27, 28, 29, 30].

FUTURE ENHANCEMENTS

CLIPS is a continually evolving product. The next release of CLIPS, version 5.0, is scheduled for release in January 1990 [31]. The primary addition to CLIPS 5.0 will be the CLIPS Object-Oriented Language (COOL). The major capabilities of COOL are: class definitions with multiple inheritance and no restrictions on the number, types, or cardinality of slots; message passing which allows procedural code bundled with an object to be executed; and query functions which allow groups of instances to be examined and manipulated. In addition to COOL, numerous other enhancements have been added to CLIPS including: generic functions (which allow different pieces of procedural code to be executed depending upon the types or classes of the arguments), integer and double precision data type support, multiple conflict resolution strategies, global variables, logical dependencies, type checking on facts, full ANSI compiler support, and incremental reset for rules.

CONCLUSION

Because of its portability, extensibility, capabilities, and low-cost, CLIPS has received widespread acceptance throughout the government, industry, and academia. The development of CLIPS has helped to improve the ability to deliver expert system technology throughout the public and private sectors for a wide range of applications and diverse computing environments.

REFERENCES

1. Culbert, C., *CLIPS Reference Manual*. NASA document JSC-22948, July 1989.

2. Riley, G., *CLIPS Architecture Manual*. NASA document JSC-23047, May 1989.

3. Giarratano, J., *CLIPS User's Guide*. NASA document, August 1989.

4. Forgy, C. "Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem". Pages 17-37, Artificial Intelligence 19, (1982).

5. Cameron, B., and Helbig, H. "Hardware Independence Checkout Software", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

6. Morris, K. "Automating Symbolic Analysis with CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

7. Hansen, R, and Flores, L. "JESNET Expert Assistant", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

8. Wannemacher, T. "Using CLIPS in a Distributed System: The Network Control Center (NCC) Expert System", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

9. Batra, S, and Macinnes, M. "GENEX: A Knowledge-Based Expert Assistant for Genbank Data Analysis", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

10. Salzman, G., Duque, R., Braylan, R., and Stewart, C. "A CLIPS Expert System for Clinical Flow Cytometry Data Analysis", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

11. Engel, B., Jones, D., Rhykerd, R., Rhykerd, L., Rhykerd Jr., C., and Rhykerd, C. "A CLIPS Expert System for Maximizing Alfalfa (Medicago Sativa L.) Production, Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

12. Thai, C., Pease, J., and Shewfelt, R. "A Decision Support System for Delivering Optimal Quality Peach & Tomato", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

13. Snyder, J., and Chirica, Laurian. "An SQL Generator for CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

14. Wess, B. "ΣCLIPSE = Presentation Management + NASA CLIPS + SQL", Proceedings of the First CLIPS Conference, Housion, TX., August 1990.

15. Charnock, E., and Eng, N. "CLIPS on the NeXT Computer", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

16. Bhatnagar, H., Krolak, P., McGee, B., and Coleman, J. "A Neural Network Simulation Package in CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

17. Orchard, R., and Diaz, A. "BB_CLIPS: Blackboard Extensions to CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

18. Geyer, S. "CLIPS Meets the Connection Machine or How to Create a Parallel Production System", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

19. Goldstein, D. "PRAIS: Distributed, Real-Time Knowledge-Based Systems Made Easy", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

20. Myers, L., Johnson, C., and Johnson, D. "MARBLE – A System for Executing Expert Systems in Parallel", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

21. Taylor, J., and Myers, L. "Executing CLIPS Expert Systems in a Distributed Environment", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

22. McGee, B., Miller, M., Krolak, P., and Barr, S. "Integrating PCLIPS into ULowell's Lincoln Logs Factory of the Future", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

23. Logie, D., and Kamil, H. "Integration of Object-Oriented Knowledge Representation with the CLIPS Rule Based System", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

24. Sobkowicz, C. "An Object Oriented Extension to CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

25. Assal, H, and Myers, L. "Implementation of a Frame-Based Representation in CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

26. Engel, B., Rewerts, C., Srinivasan, R., Rogers, J., and Jones, D. "CLIPS Interface Development Tools and Their Application", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

27. Pickering, B., and Hill, R. "HyperCLIPS: A HyperCard Interface to CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

28. Callegari, A. "Integrating Commercial off-the-shelf (COTS) Graphics and Extended Memory packages with CLIPS", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

29. Faul, B. "Constructing Complex Graphics Applications With CLIPS and the X Window System", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

30. Feagin, T. "A Graphical Interface to CLIPS Using Sunview", Proceedings of the First CLIPS Conference, Houston, TX., August 1990.

31. Riley, G., and Donnell, B. "Advanced CLIPS Capabilities", Proceedings of SOAR '90, Albuquerque, NM., June 1990. Houston

APPENDIX

CLIPS is free to NASA, USAF, and their contractors for use on NASA and USAF projects by calling the CLIPS Help Desk between the hours of 9:00 AM to 4:00 PM (CST) Monday through Friday at (713) 280-2233. Government contractors should have their contract monitor call the CLIPS Help desk to obtain CLIPS. Others may obtain CLIPS through the Computer Software Management and Information Center (COSMIC), which is the distribution point for NASA software. The program number is COS-10022. The program price is $250.00, and the documentation price is $62.00 (as of July 1990). The program price is for the source code. Price discounts are available to U.S. academic institutions. Further information can be obtained from

COSMIC
382 E. Broad St.
Athens, GA 30602
(404) 542-3265

N91

24044

UNCLAS

N91-24044

Distributed, Cooperating Knowledge-Based Systems

Walt Truszkowski
Head, Automation Technology Section
Code 522.3
Goddard Space Flight Center
Greenbelt, MD 20771

Abstract

This paper addresses some current research in the development and application of distributed, cooperating knowledge-based systems technology. The focus of the current research is the spacecraft ground operations environment. The underlying hypothesis is that, because of the increasing size, complexity, and cost of planned systems, conventional procedural approaches to the architecture of automated systems will give way to a more comprehensive knowledge-based approach. A hallmark of these future systems will be the integration of multiple knowledge-based agents which "understand" the operational goals of the system and cooperate with each other and the humans in the loop to attain the goals. The current work includes the development of a reference model for knowledge base management, the development of a formal model of cooperating knowledge-based agents, the use of a testbed for prototyping and evaluating various knowledge-based concepts, and beginning work on the establishment of an object-oriented model of an intelligent end-to-end (spacecraft to user) system. The paper will present an introductory discussion of these activities, highlight the major concepts and principles being investigated and indicate their potential use in other application domains.

Situations

Before beginning a discussion of the specific R&D activities in the area of distributed knowledge-based systems which we have been pursuing, let us take a small digression and consider the following "situations". These constitute a small sample of problems whose solutions are or will be supported in the future by access to cooperating knowledge-based systems (the computerized variety). The purpose of this initial digression is to make clear the belief that the research which is being pursued in the area of distributed cooperating knowledge-based systems has broad applicability and will eventually touch many differing aspects of human life.

1. **Distributed Knowledge Base Management:** The inputs from several distributed knowledge sources, some with access to real time sensor data, need to be collected, analyzed, checked for consistency, checked for completeness, properly annotated, and archived for easy access by researchers.

2. **Spacecraft Control:** Elements of a ground system need to be monitored and controlled in support of a space mission.

3. **Traffic Control:** There is need for an automatic rerouting and adjustment of the timing of traffic lights through a very heavy traffic zone due to an unforeseen situation.

4. **Robot Operations:** There is need for the coordination of robots which are to be engaged in carrying out a potentially hazardous experiment in a highly automated laboratory.

5. **Medical Emergency:** A team of medical experts is needed to diagnose a mysterious ailment.

6. **Utilities Control:** Life sustaining resources need to be monitored and dynamically adjusted onboard a space station in response to changing crew needs.

7. **Software Engineering:** Within the context of a knowledge-based software engineering environment agents are needed to give advise to a designer on the best match between system requirements and performance criteria and the stock of reusable/tailorable system components available.

8. **Factory Operations:** The total operations cycle of a mining plant needs to be automated for this factory being built on the moon.

9. **Electronic Library:** The facilities of a library including access to bibliographic citations, article abstracts, electronically-stored books and journal, articles, need to be made available to scholars doing research.

Each of these situations deals with a different domain, involves different activities, has different goals and objectives, addresses different problems, and utilizes different resources in accomplishing desires results. However, at an appropriate level of abstraction, automation-assisted solutions to the problems arising in these different domains can be seen to involve very sophisticated knowledge management issues and could be viewed as a tailoring of a common framework for instantiating, activating, and using teams of distributed, cooperating, knowledge-based agents. This, at least, is the hypothesis which is being formulated and being put forth for evaluation.

The rest of this paper focuses on an overview of two major research activities, i.e., the development of a reference model for knowledge base management and a formal model of cooperating of knowledge-based systems and the prototype application of some of the research results to the specific area of concern to us at the Goddard Space Flight Center, namely control center operations systems for near-earth unmanned scientific spacecraft.

Current Research Activities

Automation of control center operational systems, related to situation 2 above, is currently realized through application of single expert systems to support individual subsystem functions. State-of-the-art research in artificial intelligence and the cognitive sciences now hold that this one-to-one mapping between system function and automating agent is an inappropriate paradigm, in the extreme, which will have limited usefulness as system complexities increase. What is needed to support higher levels of automation in such systems is the use of multiple autonomous agents cooperatively providing-for and supporting desired system behaviors. This is the long-term solution to the problem of providing operational knowledge-based spacecraft control centers of the future. To support this evolution to an "intelligent" control center, our work in this area has been concentrated in the areas of knowledge base management and formal models of cooperating knowledge-based agents. A testbed for demonstrating distributed knowledge-based technologies in a spacecraft command/control environment has been established.

Knowledge Base Management

In our research, a Knowledge Base Management System (KBMS) is defined as a utility for supporting the life cycle of acquiring, refining, using, and maintaining large-scale distributed knowledge bases [Ref. 1,2,3]. The basic drivers for the KBMS research are the expectations that future autonomous systems, used in operational control centers and ground systems, will be: (1) both knowledge- and data-based, (2) distributed yet cooperating and integrated, (3) potentially large-scaled , and (4) long-lived and needing to be maintained and updated regularly. A KBMS is intended to support these drivers. Our analysis of the functional requirements for a KBMS have not been driven by general domain-independent considerations. Rather our guidance has come from what we readily expect the future (long range) control center architectures will require for operations. Some of these operational requirements are 1. a framework is needed to support the addition, connection, and operation of distributed cooperating knowledge-based systems in a rapid reliable and asynchronous fashion: 2. rules, plans, schedules, and knowledge must be continually acquired and updated from a broad array of sources with a minimum amount of human intervention; 3. knowledge verification is essential; 4. intelligent interfaces between knowledge bases and realtime sources of sensor data need to be supported; 5. a wide variety of human-factored human/machine interfaces, interaction techniques, tutoring aids, and utilities must be readily available for supporting the range of personnel which will be developing, using, or maintaining the system. To help give us an overview of the major issues to be addressed in the engineering of a comprehensive KBMS work was begun on the development of a KBMS Reference Model. In our usage of the phrase, a "reference model" refers to a map of the activities, functions to support the activities, data/information flows, interfaces/interactions which need to be supported for the general case of any KBMS. Figure 1. illustrates the current version of the KBMS Reference Model in development. The purpose of the model is to aid in identifying the major concepts associated with knowledge base management and to put these concepts in proper perspective.

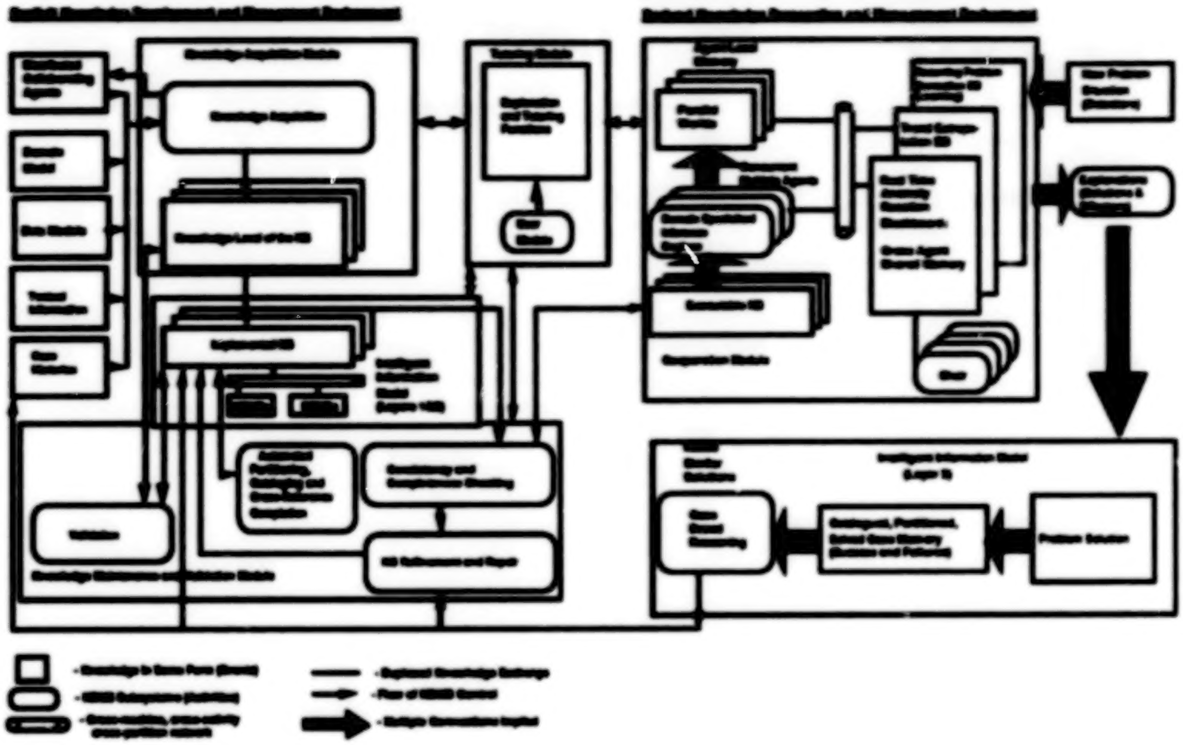


Figure 1. Reference Model for a KBMS

This reference model identifies five major functional components . These are the knowledge acquisition, knowledge maintenance and validation, tutoring, cooperation framework, and intelligent information model.

Each of these components requires a set of capabilities to carry out its role in the overall KBMS structure. An indication of these capabilities follows:

- **Knowledge Acquisition** - learn/induce, optimize, analogize, associate, specialize, generalize, experiment, ask,attach principles, annotate;

- **Knowledge Maintenance** - criticize, reward, test, trap, influence, debias, clarify, repair, refine, correct, update;

- **Tutoring** - model, display, advise, quiz, hint, explain, adapt, help, teach, monitor, oversee;

- **Cooperation** - distribute control, instantiate work breakdown structures, negotiate, broadcast, direct cast, poll, plan, dispatch, control access, propagate, search, synthesize, communicate;

- **Information Model** - cluster/abstract, subsume, access/store, sample, meta-reason, index, catalogue, model management.

This is not a complete list. Other activities may be gleaned from the reference model diagram. What is clear, even at this simplified level of description, is that the overall concept of a KBMS is a complicated and complex one involving technologies from software engineering, artificial intelligence, and the cognitive sciences.

ORIGINAL PAGE IS
OF POOR QUALITY

Work is continuing on the further refinement of the reference model. It is helping us keep a good perspective on the state-of-the-art of KBMS research. One specific problem which we are pursuing is that associated with accessing and using knowledge from heterogeneous knowledge sources. The ability to fuse knowledge from these sources into an integrated presentation to human operators is a requirement for nearer term automated ground systems operations.

Formal Modelling of Distributed Cooperating Knowledge-Based Systems

Concurrent with our work on the KBMS model we have been studying some of the mechanisms required to support cooperation among knowledge-based systems (agents) [Ref. 4,5]. Currently the use of knowledge-based components, i.e. expert systems, is rapidly becoming state-of-practice in NASA's operational ground systems. Currently ground/space network operations and the monitoring of spacecraft status data are supported in this manner. As their number and usage increases in support of operational systems it is apparent that a new system architectural concept will evolve. This new architecture will be such so as to fully exploit the computational power of these knowledge-based components and fully integrate these components into an efficient and effective operational system. The hypothesis which we share with a growing number of researchers and developers is that the new architectures will need to support distributed , cooperating knowledge-based components. As a starting point for our investigations into this area we began on the development of a logical model of cooperating knowledge-based systems. We saw the role of the model as a tool to facilitate technological research and system planning. Development of this model is continuing today. What follows is an introduction to our early work in this area.

In our analysis of what it would take to develop a team of cooperating computer-based agents to support a highly automated ground system for satellite control, several general characteristics were identified. These are:

1. There should be a logical and physical distribution of the total knowledge of the team among the agents which comprise it. **(Knowledge Partition)**

2. Each agent should have some internal model of some of the other agents in the team. **(Agent Awareness)**

3. Cooperation among agents requires communication among agents. **(Inter-agent Communication)**

4. Proper interpretation of information shared between agents may require that the agent receiving the information have access to the information context in the sender agent. **(Shared Contexts)**

5. The cooperating agents should be able to partition the problem to be solved. **(Problem Partitioning)**

6. The agents should not only be able to communicate, they should also be able to coordinate the team's activity. **(Agent Coordination)**

7. The agents in the team and the team as a whole should be capable of adjusting performance in response to environmental changes. **(Performance Adjustment)**

8. A mechanism should exist for integrating new knowledge-based agents into the team. **(Integration)**

These general characteristics will be briefly discussed in the context of the ultimate paradigm of cooperative activity among knowledge-based agents: a team of humans jointly working on a problem. This is done to help clarify the ideas and to help contribute to their justification.

Knowledge Partition: There are several reasons for supporting the partitioning of knowledge. These include the observations that partitioning supports: 1. increased performance through parallelism, 2. enhanced extensibility and maintainability of the knowledge base, and 3. well defined mappings of the knowledge to the various internal knowledge structures. When forming a team of humans it is usually desirable that each bring a special knowledge to the group and not merely know what everyone else knows. We note that knowledge overlap is however necessary to support meaningful communication.

Agent Awareness: In order to work together agents must be aware of each other's existence and attributes, i.e., some model of some of the other agents. The working of a group of humans is greatly facilitated if each team member knows something of the knowledge and capabilities of the other members. For one thing it greatly facilitates the appropriate assigning of subproblems to be solved.

Inter-agent Communication: Cooperation among agents presumes some form of communication. Both synchronous and asynchronous communication capabilities are needed for effective and efficient operation of the team. A quick unscheduled message passing among human team members is quite a common occurrence in addition to regularly scheduled status reports.

Shared Contexts: Among communicating agents each agent must be able to interpret the meaning of the information it receives. In order to ensure this the agents must have the capability to ask for the context, including underlying assumptions, of a piece of information it has received from the sending agent. Each cooperating agent needs criteria for evaluating information it receives from other agents in terms of importance, certainty, and timeliness. Irrelevant, uncertain, or out of date information may distract an agent from promising lines of reasoning. Failure t sent or receive context information may cause an agent to overlook potentially important lines of reasoning. In team meetings of humans it is quite often the case that in analyzing a new piece of information the generator of the new information is asked how it came about and how should it be interpreted in the proper context.

Problem Partitioning: In order for a team of cooperating agents to work effectively, if at all, it is not only necessary for there to be some partitioning of knowledge but also the ability to partition the work to be done. As a problem is presented to the team there need to be mechanisms available to break the problem into smaller and smaller subproblems until there are good matches between subproblem size/complexity and agent capability. Mechanisms must also exist for the synthesis of partial results, from the subproblem solutions, into a solution of the original problem. The concept of breaking a task into subtasks and making assignments to various members is a normal method of operation for teams of humans.

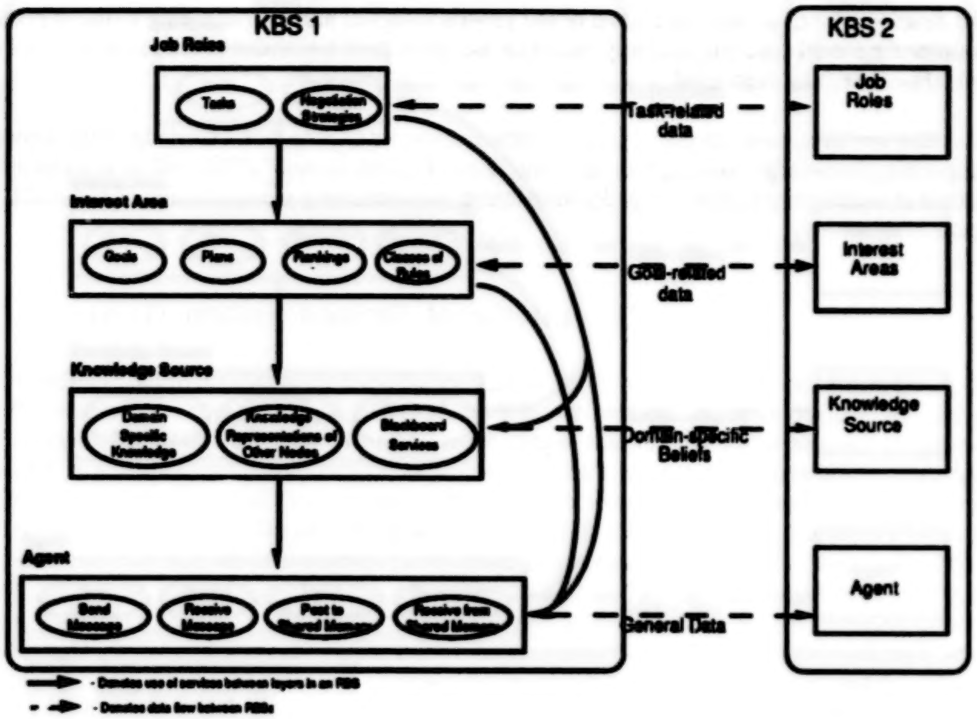
Agent Coordination: Partitioning of work ensures that each involved agent knows what it is supposed to do within the overall system. Models of other agents and communication among them ensure that the agents can interact in performing their tasks. However, synthesis of the subproblem solutions to achieve a global objective, the solution of the problem originally presented,, requires more. It requires an ability for the agents to coordinate their activities. Scheduling and control mechanisms are required to ensure the necessary degree of coordination is realized and maintained. Even very loosely-coupled teams of humans agree on the milestone schedule to be followed and may even identify one team member as the lead for a certain portion of the work.

Performance Adjustment: Depending on the nature of the problem presented, or the status of available resources, or the availability of expertise a team of agents may be required to modify its performance criteria in order to meet some overall performance objectives. Among human teams the discovery of unsuspected complexity in a portion of a problem may require regrouping and reassignments.

Integration: As new expertise is needed in both human and computer-based teams there needs to be a mechanism whereby new agents, or team members, are made a part of the team.

The reasonableness and operational semantics of these characteristics as they manifest themselves in our implementations are being evaluated.

The work on the formal modelling of autonomous agents is intended to assist in the planning, specification, development, and verification of control centers involving distributed cooperating knowledge-based systems. The current model describes a community of cooperating rule-based systems at four layers of increasing capability: (1) communicating agents - with no assumption of intelligence or rule-based capability, (2) belief-sharing knowledge sources - where a knowledge source is an agent specializing in a specific domain and a belief is any data arrived at through an inference process, (3) goal sharing interest areas - where the first indications of goal-directed reasoning appears, and (4) task sharing job roles - the fourth level of the model where overall system goals are decomposed into tasks which are allocated among various job roles. At this final level, cooperation is most fully achieved. The following figure gives an overview of the model as it is now conceived.


KBS 1
Job Roles
Tasks
Negotiation Strategies
Interest Area
Goals
Plans
Rankings
Classes of Rules
Knowledge Source
Domain Specific Knowledge
Knowledge Representations of Other Nodes
Blackboard Services
Agent
Send Message
Receive Message
Post to Shared Memory
Receive from Shared Memory
Task-related data
Goal-related data
Domain-specific Beliefs
General Data
KBS 2
Job Roles
Interest Areas
Knowledge Source
Agent
- Denotes use of services between layers in an RBS
- Denotes data flow between RBSs

The CRBS Logical Model identifies: layers of abstraction needed in KBSs to support cooperation, services and information available at each layer, and information flow between layers.

Figure 2. Logical Model

Implementation of this model is proceeding. The implementation is in C^{++} on a Macintosh II computer. An adaptation of the Contract Net protocol is being used to support the implementation. A network of a number of agents, only limited by the amount of finite memory, can be simulated in this proof-of-concept development. In the current implementation agent behaviors are determined by the protocol defined by the logical model, the Contract Net protocol, and by scripts read in and executed by the agents themselves. Various functional capabilities are being currently implemented at each of the four levels of the logical model [Ref. 6]. These are as follows:

- Level 4 - identification and representation of the dynamic attributes of other agents including their capability and availability
  - negotiated and non-negotiated assignment of work across the network based on each agent's capabilities (functional decomposition of tasks) and availability (load balancing)
  - establishing a basis for fault tolerance through a policing of contractual obligations between agents and through flexibility of inter-agent associations
  - establishing a basis for fault tolerance through flexible and dynamic association of functionality to hardware

- Level 3 - a canonical representation of goals, plans, and priorities of an agent to the outside community as a method of asserting passive influence on the behavior of outside agents
  - channels and protocols for actively influencing the goals, plans and priorities of other agents
  - evaluation of proposed work and of bids to accomplish proposed work to aid in the optimal
  - assignment of tasks to agents

- Level 2 - creation of a set of input and output daemons which accomplish the trading of beliefs between agents without the need to modify the agent's knowledge base

- Level 1 - network communication services and testing routines.

ORIGINAL PAGE IS
OF POOR QUALITY

The development of an implementation of the proposed logical model is affording us the opportunity to critically evaluate the model concepts, to firmly establish the operational semantics which the model embodies, and to change and refine the model as required.

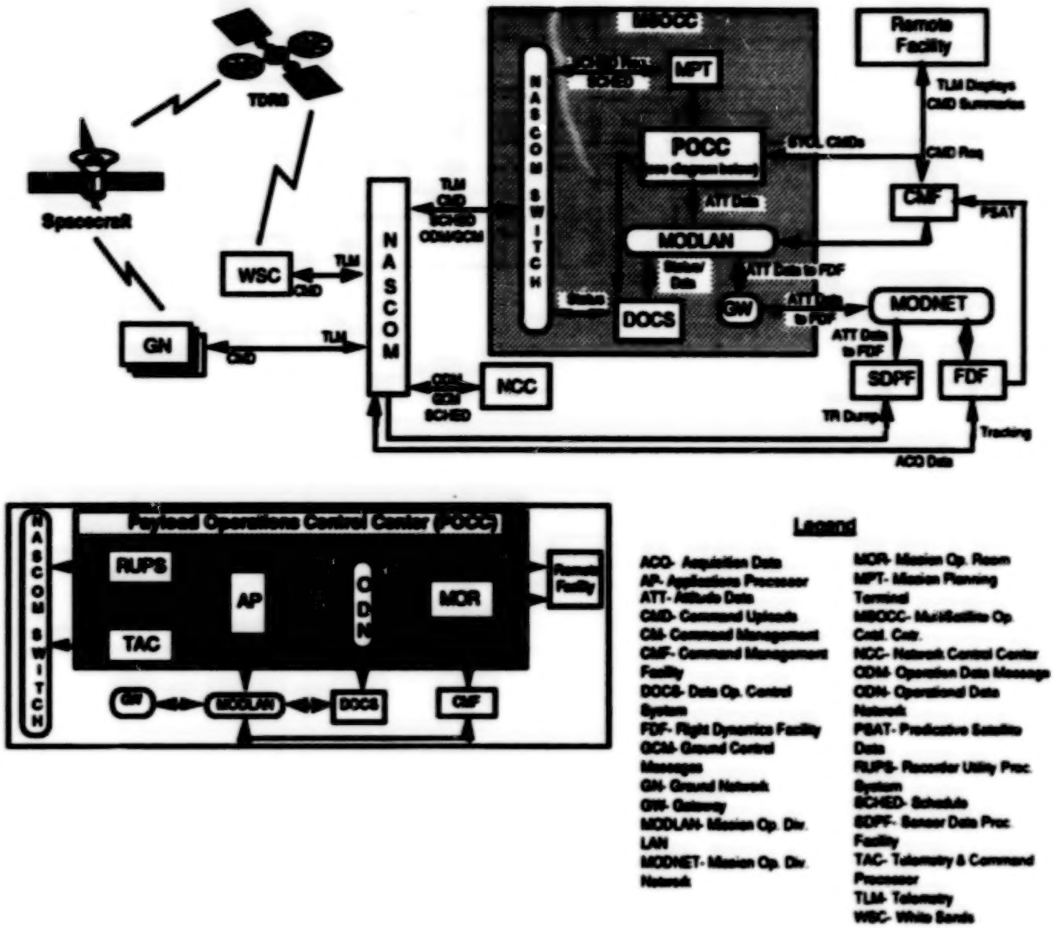
Now that you have been introduced to two research activities dealing with knowledge base management and cooperating knowledge-based agents the "Intelligent Ground System" (IGS) will be introduced. The IGS is the testbed in which we are prototyping, demonstrating, and evaluating the application of our research concepts in an operational setting.

The Intelligent Ground System (IGS)

As stated previously the major goal of our work is to influence the evolution of the systems which support the operations of space-related missions. Current ground operations systems are very complex. Though aspects of them are highly automated they nevertheless require many manually intensive operations. In some instances the cognitive workload on the human operators is reaching critical limits and the bandwidth of data and information needed to be processed by the operators is far exceeding human capability. The engineering of some of the human operator's expertise into the computer systems, i.e. the development of knowledge-based components, seems to be a viable alternative for the implementation and operation of systems with the real potential of outstripping human capability to manage. The intent of the IGS effort is to help crystalize and clarify what a highly knowledge-based ground system of the future could be like. Work on the development of the IGS is proceeding along two lines. First, a preliminary testbed, the Intelligent Control Center (ICC), has been developed. This testbed, to be discussed shortly, addresses a major component of the ground system. Experience in developing and demonstrating the ICC testbed has helped us to determine the best way to define and develop the IGS testbed. As the ICC is expanded and refined to support a higher-fidelity simulation of a portion of ground system operations work will begin on the development of an object-oriented model of the IGS. The intent is to have the ICC transform into the IGS with restructuring and the addition of several more knowledge-based components. The next sections of this paper give an overview of the ground system configuration, the portion of the ground system on which we have been concentrating our testbed efforts, and a brief overview of our current testbed configuration.

Our Current Focus

Before proceeding with a discussion of how we are focusing our research activities in a testbed environment let us briefly examine the operational domain which is intended to host the results of our work. We have been investigating the nature and effective use of cooperating knowledge-based systems in the context of ground systems needed to support the successful operation of near-earth unmanned scientific spacecraft (situation 2 above). The following Figure 3. gives a high level overview of the existing ground system.



MSOCC
Remote Facility
TLM Displays
CMD Summaries
TDRS
SCHED Req
SCHED
MPT
POCC
(see diagram below)
SYOL CMDs
CMD Req
Spacecraft
NASCOM SWITCH
TLM
CMD
SCHED
ODM/GCM
ATT Data
CMF
PSAT
MODLAN
WSC
TLM
CMD
NASCOM
Status/Data
ATT Data to FDF
Status
DOCS
GW
ATT Data to FDF
MODNET
GN
TLM
CMD
ATT Data to FDF
ODN
GCM
SCHED
NCC
SDPF
FDF
TR Dumps
Tracking
ACQ Data
Payload Operations Control Center (POCC)
NASCOM SWITCH
RUPS
AP
ODN
MOR
Remote Facility
TAC
GW
MODLAN
DOCS
CMF
Legend
ACQ- Acquisition Data
AP- Applications Processor
ATT- Attitude Data
CMD- Command Uploads
CM- Command Management
CMF- Command Management Facility
DOCS- Data Op. Control System
FDF- Flight Dynamics Facility
GCM- Ground Control Messages
GN- Ground Network
GW- Gateway
MODLAN- Mission Op. Div. LAN
MODNET- Mission Op. Div. Network
MOR- Mission Op. Room
MPT- Mission Planning Terminal
MSOCC- MultiSatellite Op. Cntl. Cntr.
NCC- Network Control Center
ODM- Operation Data Message
ODN- Operational Data Network
PSAT- Predictive Satellite Data
RUPS- Recorder Utility Proc. System
SCHED- Schedule
SDPF- Sensor Data Proc. Facility
TAC- Telemetry & Command Processor
TLM- Telemetry
WSC- White Sands

Figure 3. Functional View of the NASA End-to End Ground System

The insert box detailing the functional components in the Payload Operations Control Center (POCC) indicates that portion of the overall ground system that has been our focus for an initial analysis. Within that component we have simulated portions of the Mission Operations Room (MOR) to aid in demonstrating the use of distributed knowledge-based systems to support operations.

Intelligent Control Center (ICC) Testbed

Results from our two major research activities have been combined to support the development and operation of a distributed knowledge base testbed. This testbed, depicted in Figure 4, currently incorporates three expert systems, a spacecraft simulator, and a user interface module. This initial testbed is designed to demonstrate and test out some preliminary ideas and concepts that derive from the KBMS and modelling work and which are felt to be important for supporting advanced knowledge-based automation in future control centers. These include: communication among agents, information fusion, knowledge acquisition and refinement, information synthesis and presentation to external agents, model-based reasoning, and various levels of cooperative activity. This initial testbed is currently running in the Data Systems Technology Lab/520 at the NASA/Goddard Space Flight Center.

ORIGINAL PAGE IS
OF POOR QUALITY

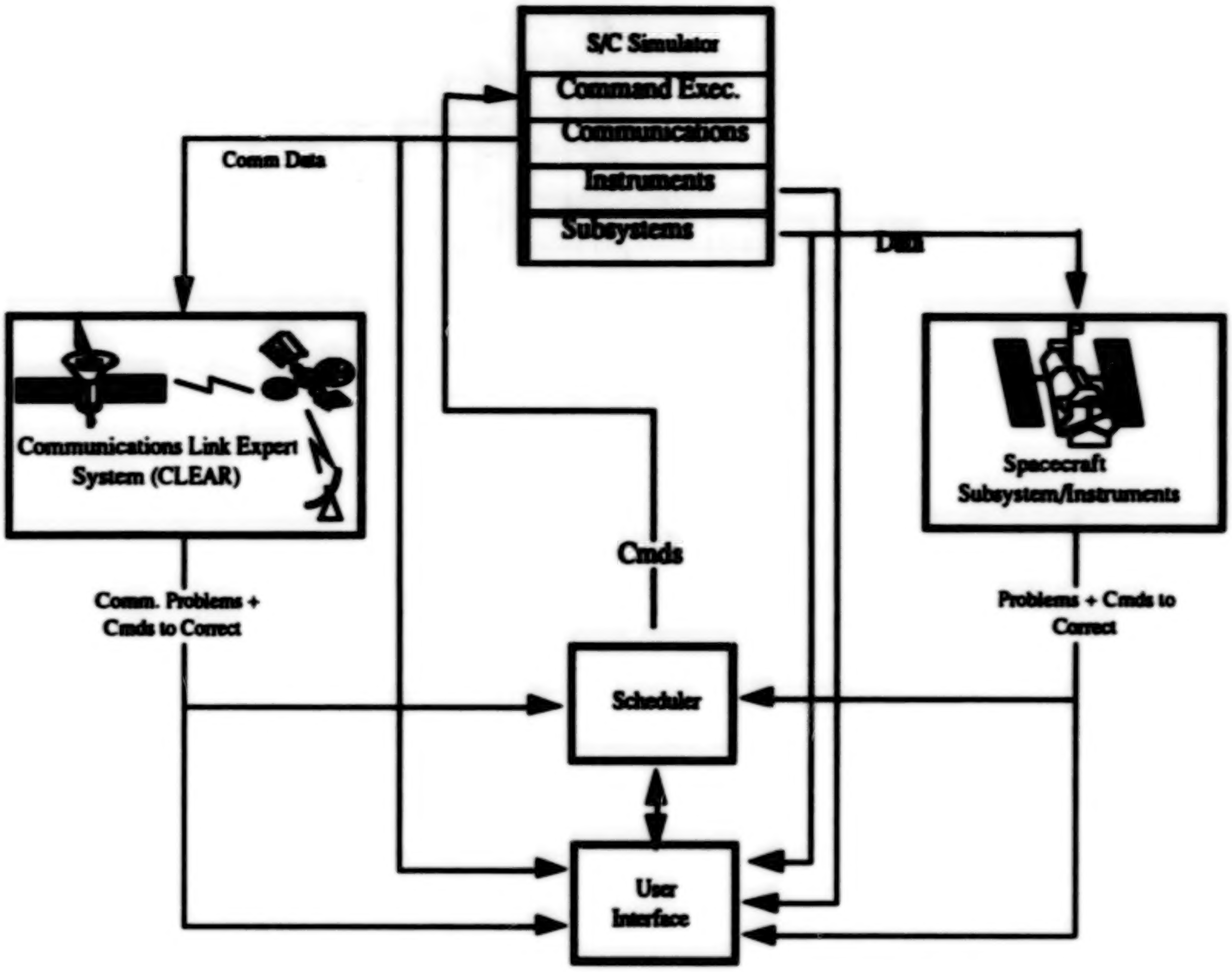

S/C Simulator
Command Exec.
Communications
Instruments
Subsystems
Comm Data
Data
Communications Link Expert System (CLEAR)
Spacecraft Subsystem/Instruments
Cmds
Comm. Problems + Cmds to Correct
Problems + Cmds to Correct
Scheduler
User Interface

Figure 4. Current Intelligent Control Center (ICC) Testbed Configuration

The purpose of this initial prototype is to demonstrate the general concept of cooperating knowledge-based systems and specifically to demonstrate the possibility of integrating existing standalone knowledge-based systems in a cooperative framework. This particular version of the testbed only addresses the functionalities associated with levels 1 and 2 of the logical model. The following is a brief description of the testbed The testbed consists of five main components: Operator's Station (User Interface), Scheduler, Spacecraft Simulator, Communications Link Expert System (CLEAR), and the Subsystem Expert System (in this case a Power subsystem). The operator uses the operator station to issue commands to the spacecraft and monitor the spacecraft's status. The scheduler schedules the commands issued by the operator for transmission to the spacecraft. The spacecraft simulator simulates the execution of the commands by the spacecraft, and it models the dynamic performance characteristics of the spacecraft's components. CLEAR diagnoses communications problems between the spacecraft and other systems, and it recommends fixes to the problems it detects. The Power expert system monitors the spacecraft's power system telemetry to detect and diagnose power system problems. The testbed components interact via a publications/subscription mechanism. The publications for a components are the output it produces, and the subscriptions for a component are the publications it receives from other components. These terms are used to reflect the fact that the information flows in the testbed are not pre-defined.

Conclusion

To date a great deal of progress has been made in understanding the role that knowledge base management and cooperating knowledge-based agents will play in the ground systems of the future. Though the level of cooperative activity and knowledge base management that has been actually achieved in the current ICC testbed is minimal, it is a good start. The expansion of the ICC to the IGS will provide a more fertile environment in which to prototype and evaluate the ideas which are being developed in the various modeling activities. One additional feature that will be added in the near future is the ability to monitor and analyze the operational performance and behavior of the testbed. Reconsidering the very brief digression at the start of this paper it seems reasonable that the framework in which we are working need not be specific to the ground operations environment but could be used to discuss the engineering of highly automated systems involving teams of computer-based cooperating agents for all the other situations identified. Exploring other applications of the concepts discussed in this paper will only help clarify our understanding of the potential benefits which will be realized through the use of cooperating knowledge-based systems.

Acknowledgements

This report has freely drawn on the works of a very talented group of people. Dr. Barry Silverman with his colleagues at the AI Lab at George Washington University and at Intellitek Inc. is the major architect of the knowledge base management system reference model. Dr. Sidney Bailin, Mike Moore, and Scott Henderson from CTA are chiefly responsible for the development of the logical model of cooperating agents. Robert Dominy and Troy Ames from Goddard/Code 522.3 are chiefly responsible for the development and demonstration of the ICC testbed. Dr. Mel Montemerlo from NASA Headquarters and Dr. Peter Friedland from NASA Ames Research Center have provided funds and guidance to us.

References

1. Truszkowski, W. F. , Silverman, B. G., Hexmore, H.: A Design Methodology for Knowledge Base Management Systems. Artificial Intelligence in NASA, AIAA Press 1989

2. Silverman, B.G., Truszkowski, W. F., et.al,: White paper Part I:Knowledge Base Management System Reference Model, Intellitek Technical Report, 1988

3. Silverman, B. G. : A Knowledge Base Management System. Institute for Artificial Intelligence, The George Washington University, 1989

4. Bailin, Sidney, et.al.: A Logical Model of Cooperating Rule-Based Systems. CTA , 1988

5. Bailin, Sidney, et.al.: Practical Issues in the Construction of Cooperative Rule-Based Systems CTA, 1989\

6. Henderson, S., Moore M.: Status of Formal Model Implementation. Private Communication, 1990

N91

24045

UNCLAS

N91-24045

ELECTRONIC NEUROPROCESSORS

Anil Thakoor

Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California 91109

ABSTRACT

Jet Propulsion Laboratory's Center - for Space Microelectronics Technology (CSMT) is actively pursuing research in the neural network theory, algorithms, and electronic as well as optoelectronic neural net hardware implementations, to explore their unique strengths and application potential for a variety of NASA, DoD, as well as commercial application problems, where conventional computing techniques are extremely time-consuming, cumbersome, or simply non-existent. This paper presents an overview of the JPL's electronic neural network hardware development activities and some of the striking applications of the JPL:s electronic neuroprocessors.

INTRODUCTION

You enter a crowded room and instantly recognize a familiar face at the far end. You wave, smile, and effortlessly weave your way through the people and furniture to reach your friend. You neither go in a straight line, nor follow a well calculated path with the least "cost" of traversing, but you proceed swiftly and efficiently anyway. In fact, your mind is already racing through the memories of a tennis game you played with that friend over ten years ago!

Digital computers today allow us to plan extremely complex, multiyear, deep space missions with amazing accuracy and provide us with enormous computing power; but something as simple to us as recognizing a face or "a pattern" in a cluttered background is not at all easy for even today's state-of-the-art supercomputers. Elaborate "expert" systems based on collective knowledge of many experts, arranged systematically to form rule bases, provide excellent software tools as "artificial intelligence". However, there is no convenient way yet to really "capture" in a computer the unique skills for example of a veteran fighter pilot with years of experience, such that they can be easily analyzed and transferred to others. The best advice from a maestro would be: "just watch how I do it and learn", something we may consider doing ... but a hopeless proposition for today's computer.

The emerging field of artificial neural networks(1), inspired by the functioning of a human brain, attempts to capture some of its unique abilities in learning, self-organizing, and intelligent information processing at extremely high speeds even with fuzzy inputs and ill-defined situations, to complement the powerful, high accuracy number-crunching digital machines. The secret of biological neural networks lies in their complex, massively parallel architectures. A human brain consists of over 10 billion neuron cells communicating among themselves through networks of over 100 trillion synaptic interconnections! Even though many of the intricacies of the brain functions are far from well-understood, it is recognized that massive parallelism, distributive storage, and a synchronous, analog, concurrent processing of information are some of its key attributes. A variety of architectural models and neural net algorithms have emerged during the past several years, through extensive software simulations, with a primary objective of developing better understanding of the unique capabilities of the biological machines and a secondary goal of capturing some of those attributes in computer systems. However, the potential of high speed from the massively parallel processing by "artificial" neural nets, mimicking the architectures of their biological counterparts, will be realized only when the architectures are actually implemented in parallel hardware(2). Can it be done? How can one build large arrays of artificial neurons and synapses and orchestrate their simultaneous operations without a system clock, as the nature does it? What are the best suited technologies and device structures? And finally, which real-world problems could substantially benefit today from the neuro-processing approach?

JPL'S APPROACH

Over the past several years, JPL's Center for Space Microelectronics Technology has sought answers to several of these questions through multifaceted research programs on theory, algorithms, hardware implementations, and applications of artificial neural networks. In particular, JPL has pioneered the development of fully parallel analog hardware implementations of neural network architectures in electronics(5-7), to understand the dynamics of such massively parallel nonlinear systems, as well as to apply the new powerful computing paradigms to problems not solved easily by other techniques.

The basic components of electronic neural network hardware are conceptually and functionally extremely simple; the neurons can be implemented as thresholding nonlinear amplifiers, and the synapses as variable resistive connections between them(8). An artificial neural network therefore consists of many simple processing elements or tailored amplifiers, representing neurons, which interact among themselves through networks of weighted connections functioning as synapses. The computation performed by the network is primarily determined by the network topography and the synaptic weights. The state of the system is identified by the pattern of activity of the neurons. Given an initial activity pattern, each neuron receives input signals from other neurons and adjusts its output accordingly over time. The system rapidly evolves into a steady activity pattern which is then interpreted as a memory recall or as a solution to a problem.

Several promising neural network architectures developed over the years utilize two broad classes of connection schemes: a fully-connected feedback architecture and a multi-layered feedforward structure, illustrated in Fig. l(a) and (b), respectively. In a feedforward network, neurons from each layer broadcast their outputs only to the neurons of the subsequent layer, modulated by the synaptic weights. The information processing thus progresses in the forward direction. In a fully connected feedback network on the other hand, all the neurons can interact dynamically with one another in parallel through the feedback synapses. Thus, the dynamics of feedback networks plays an especially important role in dictating their emergent computational properties.

The most important operational characteristic of such architectures, however, remains their massive parallelism. The highly distributed and effectively redundant information collection, storage, and manipulation in the multitude of synaptic weights give rise to inherent fault tolerance resulting in graceful degradation in their performance. On the other hand, the concurrent analog processing by a large number of neurons when implemented in hardware promises computing speeds orders of magnitude higher than serial processors. Above all, the highly parallel neural network algorithms provide unique abilities to solve computation-intensive global optimization problems and to "learn" fuzzy transformations from examples in ill-defined situations.

To be able to fabricate a variety of neural network architectures in a fully parallel fashion by using only a few selected, generic, "building blocks", JPL has developed, designed, and fabricated two separate families of cascadable custom-VLSI chips in analog CMOS: two dimensional arrays of fully programmable synapses, and one dimensional arrays of non-linear neurons.

HARDWARE IMPLEMENTATIONS

JPL's reconfigurable building blocks include a spectrum of cascadable, programmable, synaptic and multi-neuron chips (Fig. 2) with tailored functional characteristics. A typical synaptic chip consists of a fully connected 32 X 32 array of synaptic devices fabricated using the standard 2 micron bulk CMOS process. The synaptic connection embodiments vary from a simple binary (ON/OFF) scheme to fully parallel analog designs exceeding 10-bit dynamic resolution. A variety of methods have been used to obtain variable synaptic weights on the VLSI chips. For example, simple long-channel CMOS transistors provide programmable synaptic weights with binary values, "on-chip" static memories and multiplying digital to analog convertors (MDAC) furnish synapses with weight resolution of up to 7 bits, and four-quadrant multipliers that scale connection strengths according to voltages residing on invisibly charge-refreshed capacitors result in weight resolution exceeding 10 bits, a major achievement, particularly important for our currently ongoing investigations on supervised and unsupervised learning in neural network hardware. Our cascadable neuron array chips (Fig. 2) provide a unique "variable gain" feature that is valuable for embodiments of networks of varying sizes, as well as to generate controlled "simulated" annealing required during convergence of feedback networks. Such fully programmable building blocks provide a very convenient library of hardware to construct suitable network architectures dictated by a problem.

Compact hardware implementations of such massively parallel architectures of course differ significantly from conventional digital designs. For example, due to the overall power dissipation concerns in the parallel circuitry, the synaptic connections modulating communications among the neurons need to be extremely "weak" or highly resistive. Such unusual requirements of the massively parallel and nonlinear processing in neural net hardware present a totally new set of interesting issues regarding precision and tolerance, influence of static and dynamic noise sources, and the useful dynamic resolution in the analog information being processed. The synaptic connection elements on a CMOS chip for example utilize high-precision, long channel field effect transistors, providing weak, current-limiting connections in their fully "ON" state and several orders of magnitude higher resistance in the "OFF" state. Furthermore, resistivity-tailored thin film elements of cermets integrated with the memory-switching devices promise high density ($\sim 10^9$ synapses/cm^2) for the synaptic arrays.

NOVEL DEVICE STRUCTURES

Although conventional VLSI technology offers a convenient approach to implementing "neuro-functions" in hardware, its circuit concepts tailored for sequential processing result in unnecessary hardware complexity to accomplish simple functions. For example, a programmable analog synaptic connection, essentially a resistive component, requires a large number of transistors occupying expensive silicon area and thus limits the size of the implemented network. Ideally, a two-terminal programmable solid state memory device would simplify large network implementations significantly. For the next generation neural network hardware therefore, JPL is investigating several novel thin film device structures based on materials with tailored electronic properties and analog thin film-VLSI hybrid device concepts. JPL's thin film device efforts have already demonstrated programmable nonvolatile synapses based on memory switching in hydrogenated amorphous silicon and manganese oxide with a potential to realize extremely high synaptic density, approaching 10^9 synapses/cm^2, suited for a variety of massive data management applications. Furthermore, optically addressable, analog memory devices based on ferroelectric thin films are also under development for large, 2-dimensional (focal plane) synaptic arrays for optoelectronic implementations of neural networks.

NEUROPROCESSOR APPLICATIONS

Unique strengths of neural networks complement the power of conventional digital computers very well. Hardware implementations of tailored neural net architectures therefore become extremely high speed, special-purpose, function-specific "co-processors" interfaced to digital computers. JPL is heavily involved in developing such application-specific neuro-processors for complex problems, where digital techniques are either limited in scope and speed or simply not applicable due to the computation-intensive and fuzzy nature of the problems. Figure 3 shows an example of a VLSI neuroprocessor interfaced to a personal computer. The neural network in this case is processing several analog constraints (e. g. conditions of soil, surface roughness, slope, vegetation, and rain) simultaneously, to determine the "cross-country mobility" for a military vehicle over a terrain under consideration, with a manyfold speed enhancement compared to the digital machine alone.

Another striking example of the enabling nature of the neuroprocessing approach is evident from our dedicated neuroprocessor for resource allocation, currently under development. Based on our innovative "analog prompt scheme" and "limited connectivity" architecture already demonstrated in hardware, the resource allocation processor promises real-time solutions to computation-intensive problems of global optimization and dynamic assignment such as pairing of resources to consumers (or assignment of weapons to targets) to minimize the "global cost" involved in the situation. For example, a 64 resource to 64 consumer assignment problem, for a one-to-one association, involves a cost matrix of 64 X 64, and a total of 64! or 10^{89} possible solutions. For a problem of this size, the neuroprocessor promises an optimal or near-optimal solution within a fraction of a millisecond. This is over three orders of magnitude faster compared to conventional heuristic search techniques, even when running on multiprocessor machines such as hypercube. Moreover, the reconfigurable neuroprocessor offers solutions to dynamic assignment problems even with arbitrary many-to-many association constraints, extremely difficult for digital computing methods.

Other neuroprocessors developed by JPL are under evaluation at present for applications in cartographic analysis, terrain feature recognition from landsat imagery, and best path determination in a constrained space. Furthermore, JPL has developed one of the first ever reconfigurable, multilayer neuroprocessor systems with a capability of learning in analog hardware. This system is currently applied to problems of computation-intensive inverse kinematic transformations in robotics and ill-defined feature recognition from multispectral images.

Clearly, the strength of neuroprocessors is complementary to the digital computers. To combine the best of both worlds, The future supercomputers with multiple digital processing nodes (e. g. hypercube architecture), may have several special purpose neuroprocessors "servicing" individual computing nodes. Such a system may also have additional analog neuroprocessors for example to carry out specialized tasks such as load balancing and problem decomposition, where neural network-derived methods show great promise.

CONCLUSIONS

JPL's fully parallel, programmable, neural network hardware has not only provided high speed research tools to investigate unique emergent computational properties of neural networks, but has also furnished the "building blocks" for development of special-purpose, application-specific neuroprocessors. Powerful artificial neural networks have been implemented in fully parallel hardware, in spite of the inherent peculiarities and unavoidable noise constraints, characteristic of analog hardware. Analog, parallel neuroprocessors provide orders of magnitude speed enhancement and/or totally new capabilities compared to conventional digital techniques.

ACKNOWLEDGEMENTS

The work described in this paper was performed by the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored in parts by the Defense Advanced research Projects agency, the Joint Tactical Fusion Program Office, and the Strategic Defense Initiative Organization/Innovative Science and Technology, through an agreement with the National Aeronautics and Space Administration.

REFERENCES

1. McClelland, J. L.; and Rumelhart, D. E. : Parallel Distributed Processing. Part I and II. The MITPress, (Cambridge, MA), 1987.

2. Mead, C. : Analog VLSI and Neural Systems. Addison Wesley, USA, 1989.

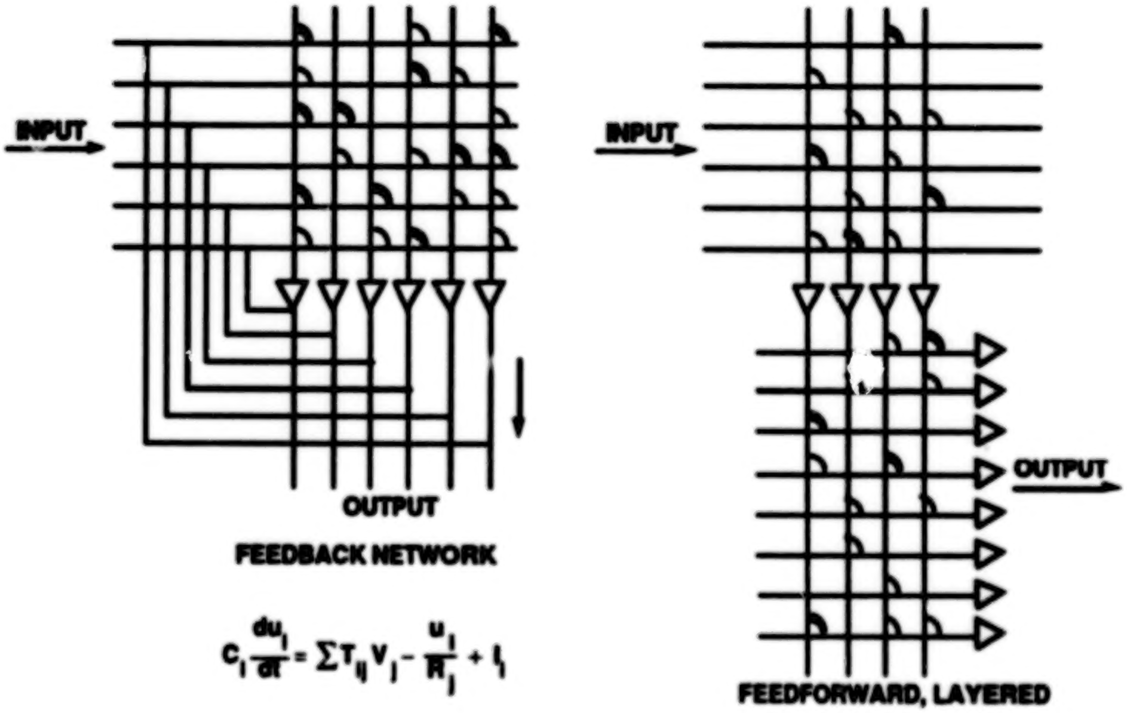
3. Thakoor, A. P.; Moopenn, A.; Lambe, J.; and Khanna, S. K. : Electronic Hardware Implementations of Neural Networks. Applied Optics, vol. 26, no. 23, 5085, 1987.

4. Eberhardt, S.; Duong, T.; and Thakoor, A. P. : Design of Parallel Hardware Neural Network Systems from Custom Analog VLSI Building Block Chips. Proc. IEEE/INNS Int'l Joint Conf. on Neural Networks, vol. 2, 183, 1989.

5. Moopenn, A.; Lambe, J.; and Thakoor, A. P. : Electronic implementation of associative memories based on neural network model. IEEE Trans. Sys., Man, and Cyber., SMC-17, 325, 1987.

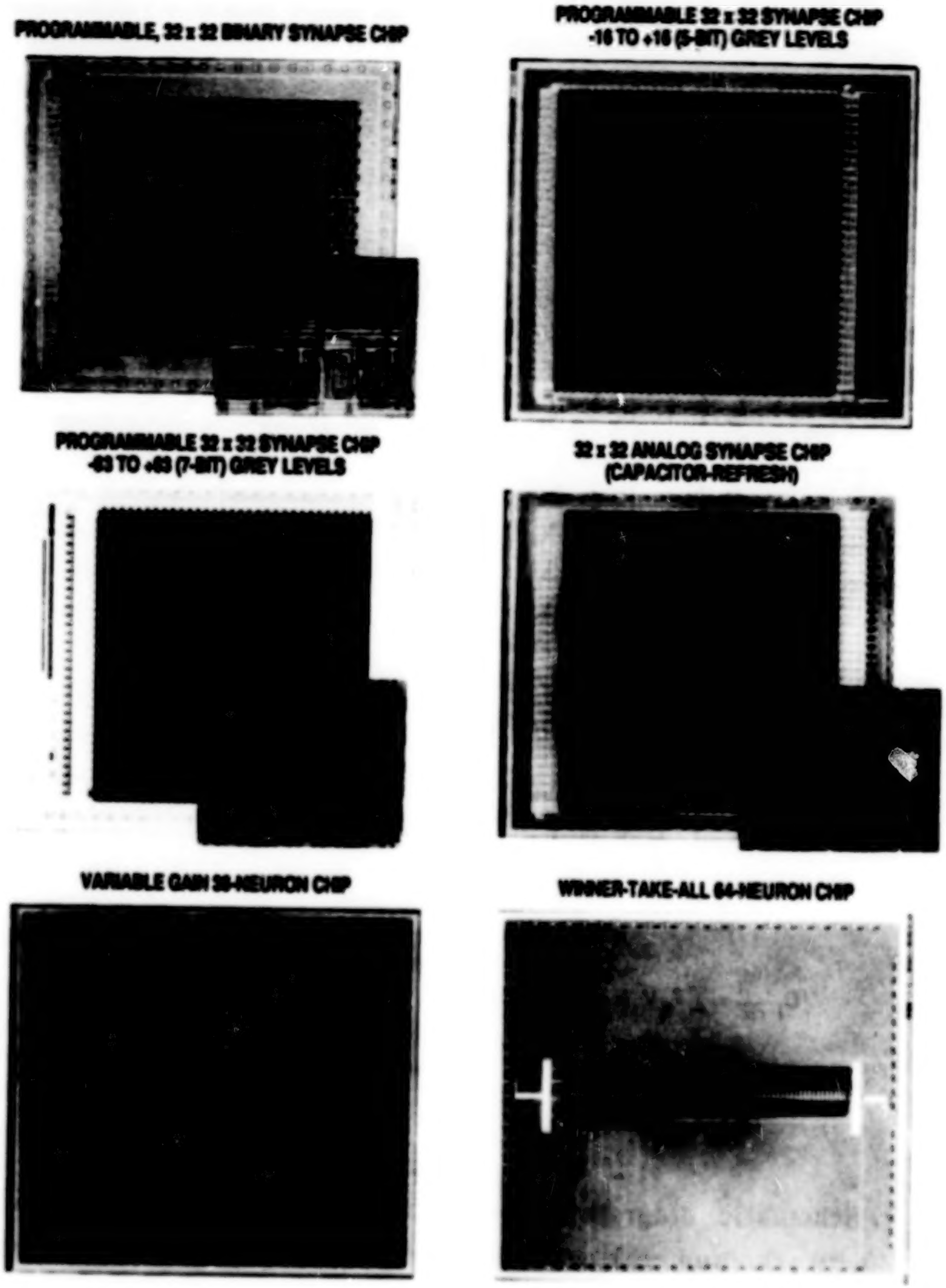
6. Daud, T.; Moopenn, A.; Lamb, J. L.; Ramesham, R.; and Thakoor, A. P. : Neural Network Based Feed-Forward High Density Associative Memory. Proc. IEDM, 107, 1987.

7. Eberhardt, S.; Duong, T.; and Thakoor, A. P. : A VLSI Analog Synapse Building Block Chip for Hardware Neural Network Implementations. Proc. Third Annual Parallel Processing Symposium, Vol.1, 257, 1989.


INPUT
OUTPUT
FEEDBACK NETWORK
$C_i \frac{du_i}{dt} = \sum T_{ij} V_j - \frac{u_i}{R_i} + I_i$
INPUT
OUTPUT
FEEDFORWARD, LAYERED

(a) (b)

Figure 1. Schematic diagram of (a) a fully connected feedback neural network and (b) a multilayer feed-forward neural network architecture.



PROGRAMMABLE, 32 x 32 BINARY SYNAPSE CHIP

PROGRAMMABLE 32 x 32 SYNAPSE CHIP
-16 TO +16 (5-BIT) GREY LEVELS

PROGRAMMABLE 32 x 32 SYNAPSE CHIP
-63 TO +63 (7-BIT) GREY LEVELS

32 x 32 ANALOG SYNAPSE CHIP
(CAPACITOR-REFRESH)

VARIABLE GAIN 36-NEURON CHIP

WINNER-TAKE-ALL 64-NEURON CHIP

Figure 2. Cascadable custom-VLSI neural network building block chips fabricated using analog CMOS technology.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE IS
OF POOR QUALITY

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

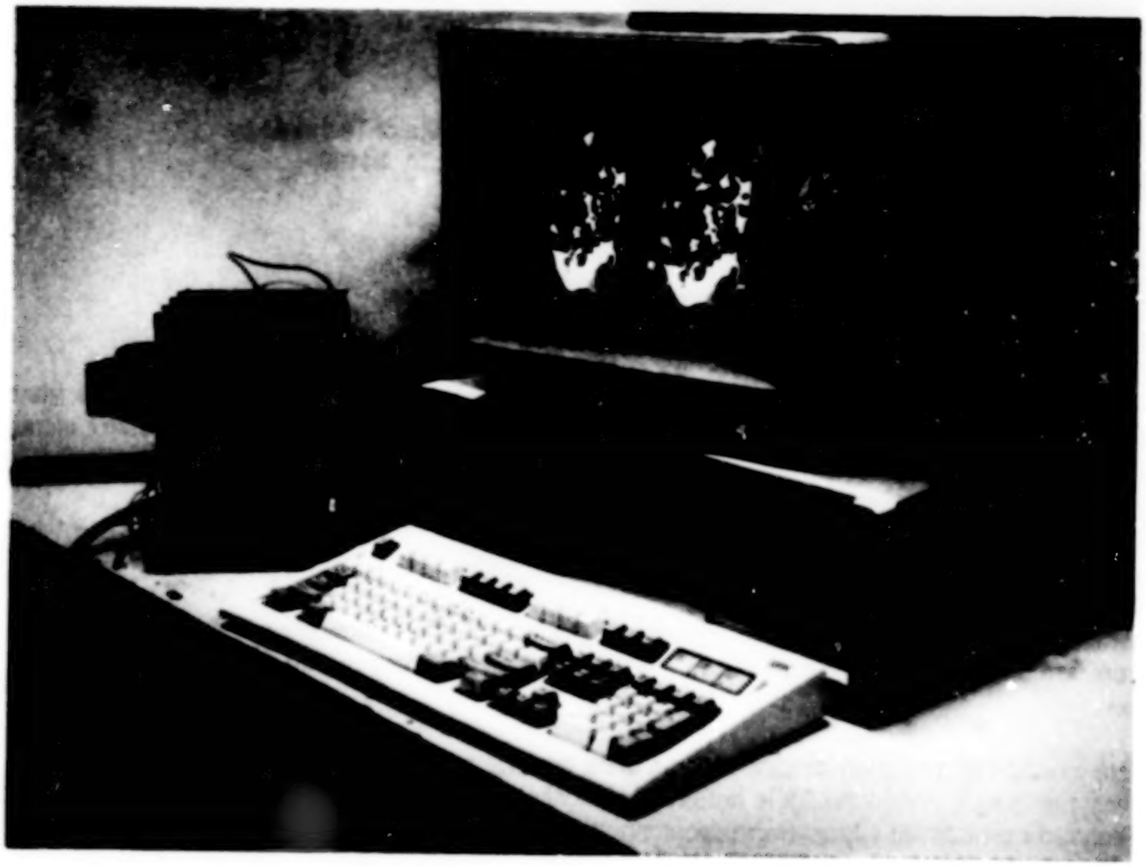


Figure 3. A neuroprocessor interfaced to a digital computer, configured for terrain trafficability determination.

ORIGINAL PAGE IS
OF POOR QUALITY

N91

24046

UNCLAS

N91-24046

Intelligent Vision System for Autonomous Vehicle Operations

Marija S. Scholl

Jet Propulsion Laboratory, California Institute of Technology
4800 Oak Grove Drive, Pasadena, CA 91109

ABSTRACT

We describe a complex optical system consisting of a 4f optical correlator with programmable filters under control of a digital on-board computer that operates at video rates for filter generation, storage, and management.

ROBOTIC PLANETARY EXPLORATIONS

Exploration of an unknown environment has traditionally required human presence to classify the environment and the objects within it, and to make survival decisions on the basis of input to the senses.

Planetary exploration to such destinations as Mars and the Moon, however, is to an environment unfriendly to human habitation. It is not only physically demanding, it also represents an operational challenge to maintain life support and protection systems. The obvious advantage of using an intelligent machine for exploration is protection of human life in a potentially adverse environment.[1] Additionally, a machine is self-sufficient in an environment that does not support biological life. An artist's conception of a roving vehicle exploring the Martian surface is shown in Figure 1.

A semi-autonomous vehicle with locomotion, visual and tactile sensors, and some on-board intelligence is capable of surveying large surface areas to locate suitable landing, habitat, and operations center sites, and to perform preliminary geological investigations prior to human exploration. This exploratory vehicle may be equipped with a combined vision–image classification system based on optical correlator technology.[2] The application of the correlator to the object-recognition task is illustrated in Figure 2.

OPTICAL CORRELATOR SYSTEM

The optical correlator system consists of four modules: the scene coherent-light conversion module; the optical correlator bench; the matched filter module; and a stand-alone on-board digital computer for filter storage, management, and on-board filter generation. Figure 3 is a photograph of the optical correlator system. Figure 4 shows the two personal computers that are used to operate the correlator system. The first one is used for filter management and control. The other is dedicated to the tracking function of the correlator.

Capabilities

Using the optical correlator system to navigate the semi-autonomous vehicle on an uncharted planetary surface requires the following capabilities:

a) Recognition of objects (object clusters and features)
b) Recognition of rotated and magnified objects
c) Recognition of objects upon change of perspective
d) In-situ generation of reference filters

Optical correlator system

The optical correlator performs the basic function of recognizing an object when the matched filter has been generated with prior knowledge. The 4f configuration incorporated in this system, shown in Figure 5, is based on the classical Vander Lugt correlator.[3] An actual photograph of the correlator optical system is shown in Figure 6. It was designed as a laboratory prototype instrument for ease of adjustment and as a test bed for further optimization. More compact optical correlators can be built once the technology has been demonstrated in the laboratory environment.[4]

Scene input is presented in coherent light to the Fourier transform lens group, consisting of a positive Fourier transform (FT) lens and a negative Fourier transform Barlow lens. M3 is a beam-turning mirror which, like all the other mirrors, fits the optical correlator on a standard 24- by 48-inch optical table.

The Fourier transform of the scene is obtained on the surface of the liquid crystal light valve after the polarized beam passes through the polarizing beam-splitter cube. A programmable video display is used to project the matched filter on the liquid crystal light valve.

The liquid crystal light valve in the filter module reflects at those pixel locations where a video display and the scene Fourier transform have a bright pixel at the same time. The reflected light is polarized, so the beam splitter reflects it into the correlation arm. An imaging correlation lens images the surface of the liquid crystal light valve on the CCD camera after two more beam-turning mirrors. A binarizing polarizer is used to increase contrast between dark and light pixels.

Incoherent light from the scene is translated into coherent light in the input arm of the optical correlator by using a video display—liquid crystal light valve combination with the HeNe laser beam as the read beam. A video camera provides video input to the display, which is coupled to the liquid crystal light valve using a fiber optics faceplate. A properly conditioned laser beam, after passing through a spatial filter for clean-up, an aperture for beam limiting, and a collimating lens, uniformly illuminates the spatial light modulator. The input scene in coherent and polarized light is transmitted through the beam splitter and fed into the optical correlator.

Rotation and magnification

A Vander Lugt optical correlator recognizes an object and finds its location in the same orientation and same size as the object used for generating the matched filter. In this correlator system, a CRT display is used for addressing the liquid crystal spatial light modulator. Thus, the rotated and magnified object can be recognized by derotating and demagnifying the input scene on the CRT raster using auxiliary electronics.

Perspective change

Perspective change for a slowly moving exploratory vehicle is not expected to be appreciable. It will be treated as a linear combination of a small magnification change as the vehicle approaches the object and a small rotational change as the vehicle moves past the object.

In-situ filter generation

The video camera may be used to capture the image of an object under new conditions, such as an appreciable change in illumination or a change in perspective. The scene is stored on the frame grabber. The object may be isolated and a new matched filter generated using the on-board computer. The object coordinates can be recalled from the on-board computer or relayed to the mobile vehicle by a remote human operator.

Filter management and control

The optical correlator system is flexible in recognizing objects because it uses binary phase-only filters[5] as matched filters. Thus, only zero or one is stored for each pixel position, requiring only 525 x 525 memory locations. Only the central 33 percent of the filter frame is used, because the higher frequencies have not been found to contain much information. This further reduces memory requirements. Such small memory usage, combined with video rates of filter recall, results in a correlator that possesses a large repertoire of matched filters that can be recalled at video rates.

APPLICATIONS TO PLANETARY EXPLORATIONS

Figure 7 shows a hypothetical Martian surface as seen by a camera on board the roving vehicle. The rover is pursuing a route among rocks. It has been previously instructed to make a left turn after the dark rock on the left by a communication orbiter.

The autonomous vehicle uses the dark rock (outlined) to generate a binary phase-only matched filter, as shown in Figure 8. The correlation peak is shown in Figure 9 for a 0° viewing angle. When the autonomous vehicle turns at the rock, the correlation intensity decreases. At 10° rotation to the left of normal, the peak is attenuated, as shown in Figure 10. Although the peak intensity appears the same here due to the photographic process, the noise becomes more prominent, decreasing the signal-to-noise ratio. Figure 11 shows the relative correlation peak intensity as a function of rotation angle for the dark rock. With an increase in rotation angle beyond 10 °, a decrease in the correlation intensity requires generation of a new matched filter of the dark rock.

CONCLUSION

The hybrid digital–optical cross correlator is highly suitable for image recognition and feature classification in support of semiautonomous robotic explorations.

ACKNOWLEDGMENTS

The author wishes to thank Jan Ralls of the Jet Propulsion Laboratory for encouragement and helpful discussions. The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the Research Directorate of the U.S. Army Missile Command in Redstone Arsenal, Alabama.

REFERENCES

1. M. S. Scholl, M. S. Shumate, J. A. Sloan, "Hybrid Digital–Optical Cross Correlator for Image and Feature Classification," *Optics in Complex Systems, SPIE Proceedings 1319*, 1990.

2. J. A. Sloan and L. A. Holloway, Jr., "A Self-Serving Optical Correlator for Tracking," *Optical Information Processing Systems and Architectures, SPIE Proceedings 1151*, 1989.

3. A. B. Vander Lugt, "Signal Detection by Complex Spatial Filtering," *IEEE Trans. Inform. Theory, IT-10*, 139, 1964.

4. D. A. Gregory, "Compact Optical Correlators," *Real-Time Signal Processing for Industry Applications, SPIE Proceedings 960*, 1988.

5. D. M. Cottrell, R. A. Lilly, J. A. Davis, and T. Day, *Appl. Opt. 26*, 3755, Sept. 15, 1987.

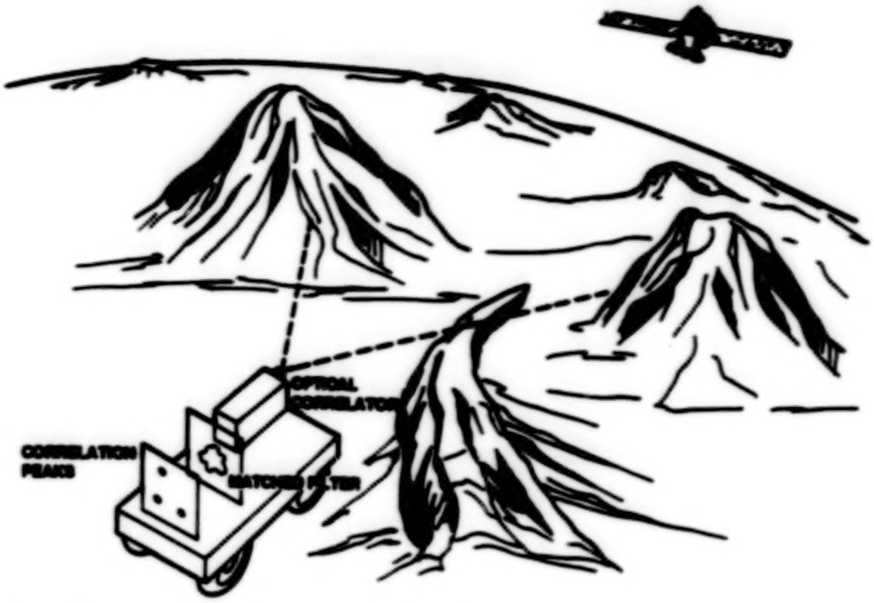

OPTICAL CORRELATOR
CORRELATION PEAKS
MATCHED FILTER

Figure 1. Artist's conception of a roving vehicle exploring the Martian surface

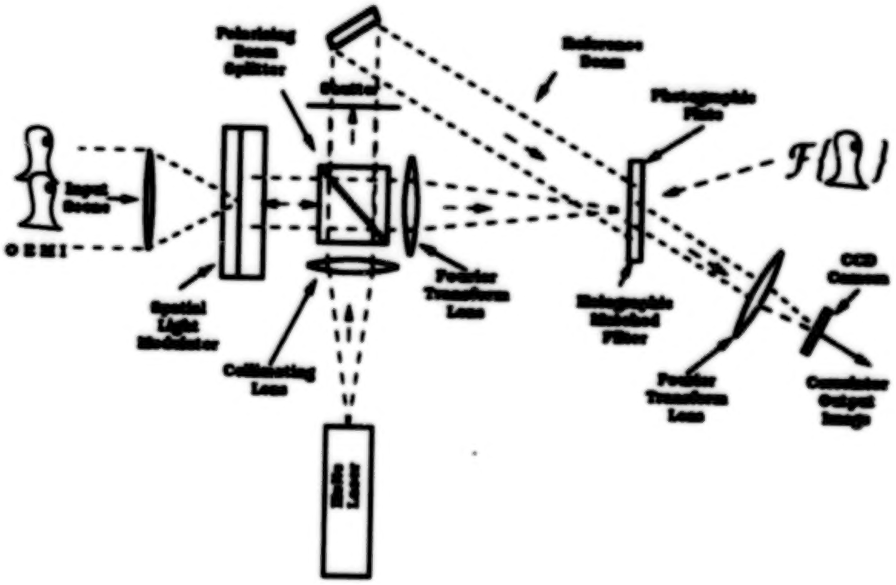

Polarizing Beam Splitter
Shutter
Reference Beam
Photographic Plate
Input Scene
Spatial Light Modulator
Collimating Lens
Fourier Transform Lens
Holographic Matched Filter
CCD Camera
Fourier Transform Lens
Correlator Output Image

Figure 2. Correlator technology applied to object recognition

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

Figure 3. Photograph of the optical correlator system

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

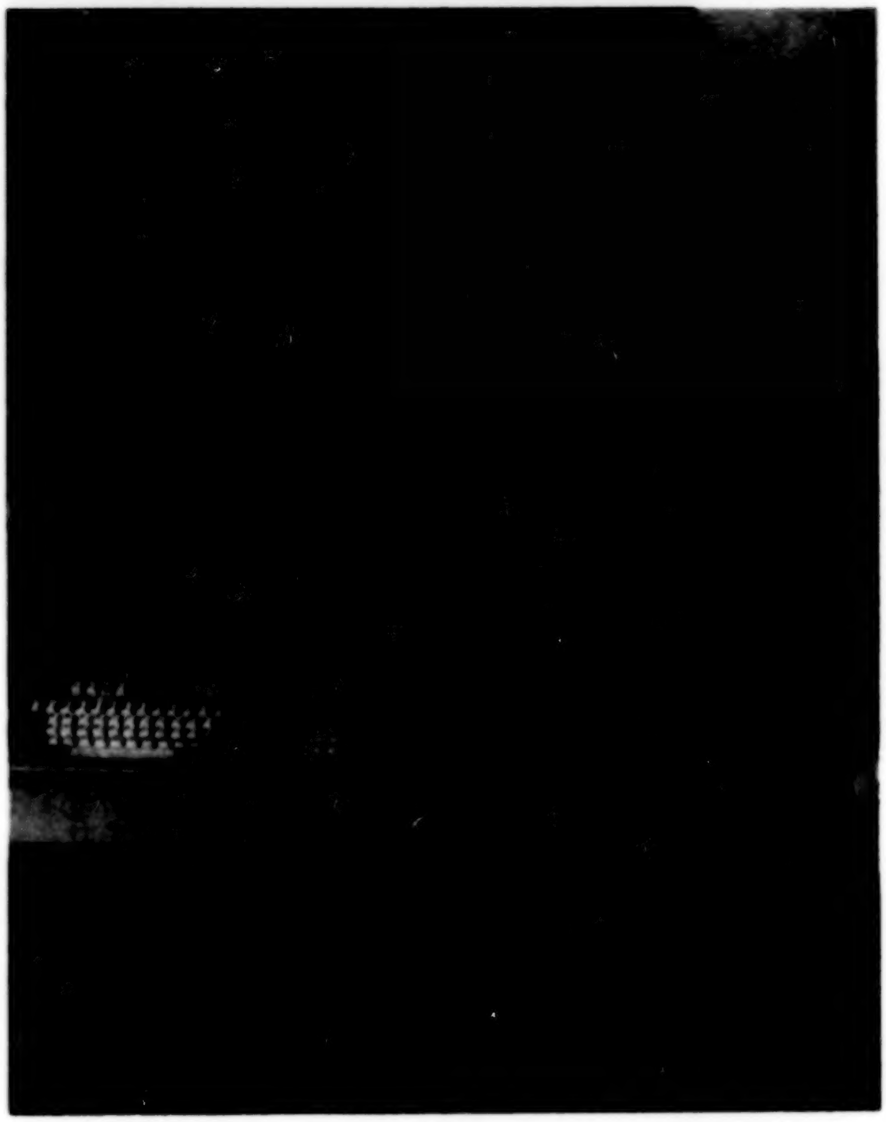


Figure 4. Photograph of the PC computers used for filter management and object tracking

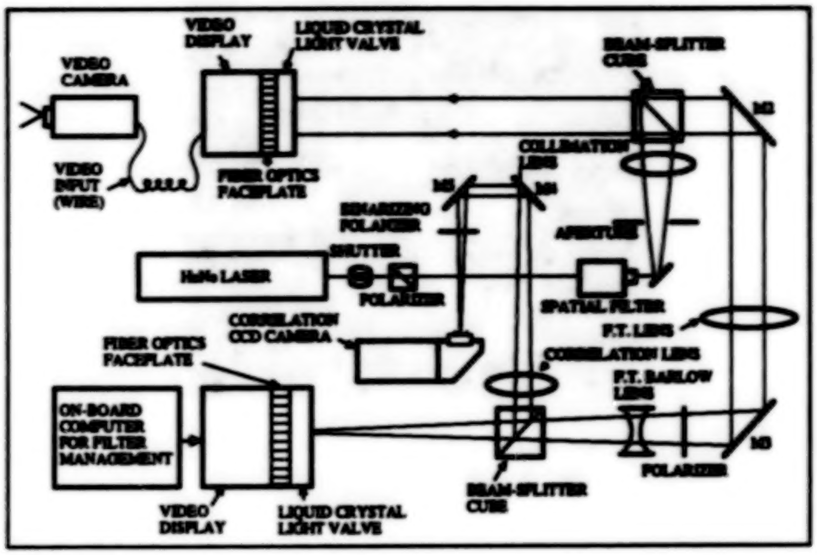

VIDEO CAMERA
VIDEO DISPLAY
LIQUID CRYSTAL LIGHT VALVE
BEAM-SPLITTER CUBE
M2
VIDEO INPUT (WIRE)
FIBER OPTICS FACEPLATE
COLLIMATION LENS
BINARIZING POLARIZER
APERTURE
SHUTTER
HeNe LASER
POLARIZER
SPATIAL FILTER
CORRELATION CCD CAMERA
F.T. LENS
FIBER OPTICS FACEPLATE
CORRELATION LENS
F.T. BARLOW LENS
ON-BOARD COMPUTER FOR FILTER MANAGEMENT
M3
POLARIZER
BEAM-SPLITTER CUBE
VIDEO DISPLAY
LIQUID CRYSTAL LIGHT VALVE

Figure 5. Schematics of the correlator optical system

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

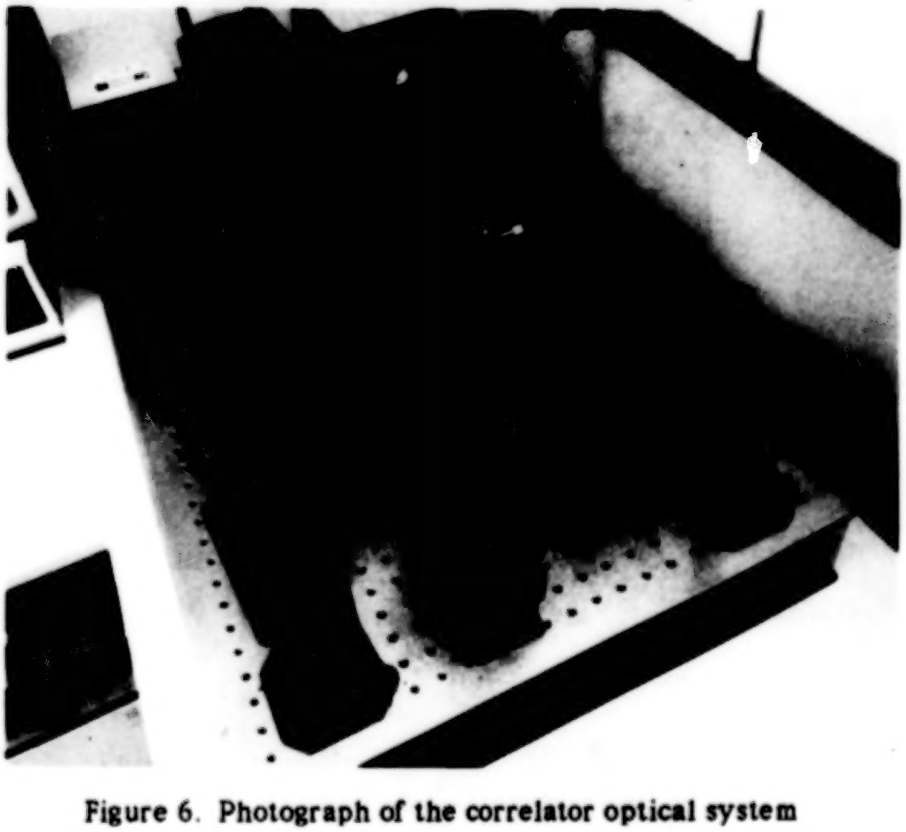


Figure 6. Photograph of the correlator optical system

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

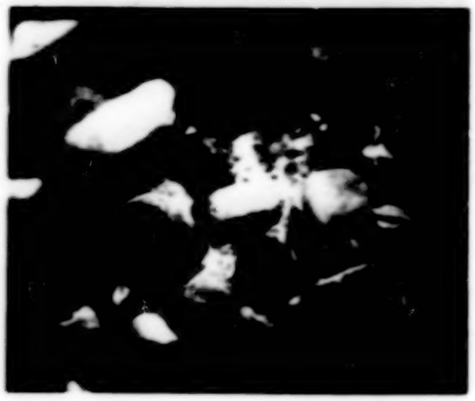


Figure 7. Hypothetical Martian surface as seen by on-board camera on the roving vehicle

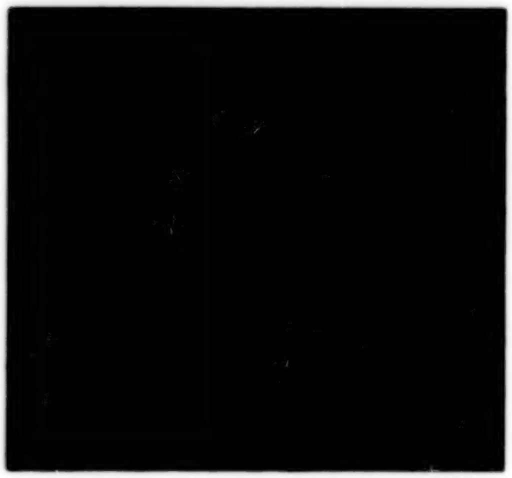


Figure 8. Binary phase-only filter of the dark rock on the left side of Figure 7

ORIGINAL PAGE IS
OF POOR QUALITY

Figure 9. Correlation peak due to recognition of dark rock - direct view

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

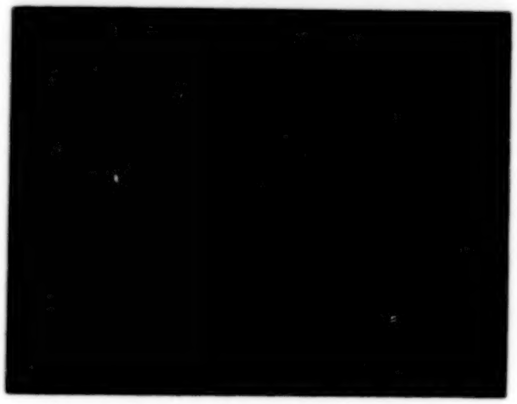


Figure 10. Correlation peak due to recognition of dark rock at- 10 degree rotation

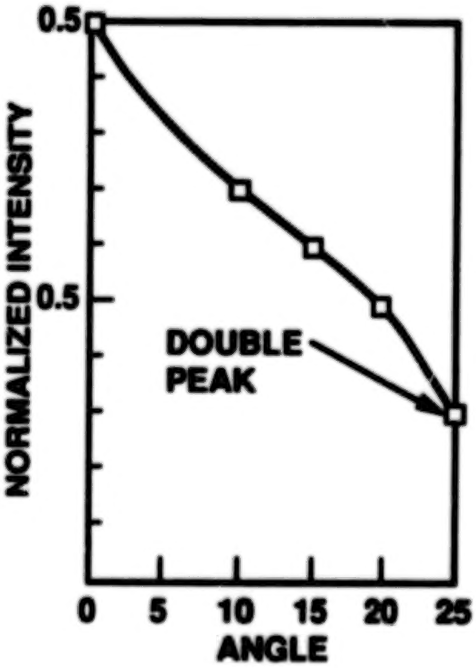


Figure 11. Relative peak intensity of the correlation peak as a function of rotation angle

ORIGINAL PAGE IS
OF POOR QUALITY

SESSION B - COMPUTER TECHNOLOGY AND SOFTWARE ENGINEERING (PART 2)

Wednesday November 28, 1990

- **Software Reengineering**
- **Transportable Applications Environment (TAE) Plus: A NASA User Interface Development And Management System**
- **Applications Of Fuzzy Logic To Control And Decision Making**
- **Genetic Algorithms**
- **Integrated Vertical Bloch Line (VBL) Memory**

PRECEDING PAGE BLANK NOT FILMED

N91

24047

UNCLAS

N91-24047

SOFTWARE REENGINEERING

Ernest M. Fridge III
Software Technology Branch/PT4
NASA/Johnson Space Center
Houston, Texas 77058
(713) 483-8109
Nasamail: EFRIDGE

INTRODUCTION

Programs in use today generally have all of the functional and information processing capabilities required to do their specified job. However, older programs usually use obsolete technology, are not integrated properly with other programs, and are difficult to maintain. Reengineering is becoming a prominent discipline as organizations try to move their systems to more modern and maintainable technologies. Johnson Space Center's (JSC) Software Technology Branch (STB) is researching and developing a system to support reengineering older FORTRAN programs into more maintainable forms that can also be more readily translated to a modern language such as FORTRAN 8x, Ada, or C. This activity has led to the development of maintenance strategies for design recovery and reengineering. These strategies include a set of standards, methodologies, and the concepts for a software environment to support design recovery and reengineering.

This document provides a brief description of the problem being addressed and the approach that is being taken by the STB toward providing an economic solution to the problem. A statement of the maintenance problems,the benefits and drawbacks of three alternative solutions, and a brief history of the STB's experience in software reengineering are followed by the STB's new FORTRAN standards, methodology, and the concepts for a software environment.

STATEMENT OF THE PROBLEM

Based on trends in the computer industry over the last few years, it is clear that computer hardware, languages, and procedures are not static. The software industry recognizes that a large existing software base must be dealt with as new software engineering concepts and software technologies emerge. The old systems use outdated technology and are costly to maintain. At JSC, as in industry at large, there is a large investment in existing FORTRAN software. These FORTRAN systems do not consistently use modern software practices that can increase maintainability. Yet these systems must be maintained for perhaps the next 20 years. Management is seeking ways to reduce maintenance costs.

In the 1960s-70s many FORTRAN programs were developed at JSC, each with its own sizeable software development team, and its own input/output format. These programs could not communicate readily and eventually were "wired" together in a very crude semblance of integration. Standards could not be enforced because FORTRAN did not enforce them and some were not visible by just looking at the code. The problem was aggravated by the lack of training of new developers plus a 50 percent turnover in the very large development staff every two years. In addition, the user organizations had more people doing development than the development group, and these other organizations were not always aware of the standards and support tools available. This history has left JSC with the following problems:

- Many programs are large and difficult to understand, resulting in maintenance problems.
- The problems in maintenance led to users keeping their own versions of programs, resulting in tremendous duplication.

Many of the FORTRAN programs have already been converted from their original dialect of FORTRAN to the FORTRAN 77 standard. Additional conversions will periodically be required even if only to new FORTRAN standards. It is necessary to consider the question, where will that code have to be in five or ten years? Three possible answers come to mind:

- FORTRAN 77 is the current standard, but the next FORTRAN standard, FORTRAN 8x, is close to release. As vendors stop supporting FORTRAN 77, existing FORTRAN will have to move to the new standard or to another language.

- Much of the code may move to the Ada language. This will be particularly true on Space Station Freedom work.

- With C being the language of choice for Unix , some of the code might move to the C language.

ALTERNATIVE SOLUTIONS

Three alternative solutions to the problems identified above have been identified: complete redevelopment of the program, code translation to a more modern language or version of a language, and reengineering. Each of these is illustrated in figure 1 and discussed briefly in the following paragraphs.

Redevelopment of a system from scratch is very expensive. Redevelopment includes all of the same phases of the life cycle as new development, from requirements through integration and testing. Extensive domain analysis is required, and there is a risk of incomplete requirements. All too often it is reported that a large program will be redeveloped from scratch to a more modern style only to find out that the new developers did not understand all of the functions and necessary information requirements of the existing system.

Code translation, especially automatic code translation, costs much less. Some might then ask, why worry about all of this now? We can use a translator when the time comes that we are forced to move the code forward. Although this would be a nice solution, the truth is that code translators have proven unsuccessful due to several major reasons:

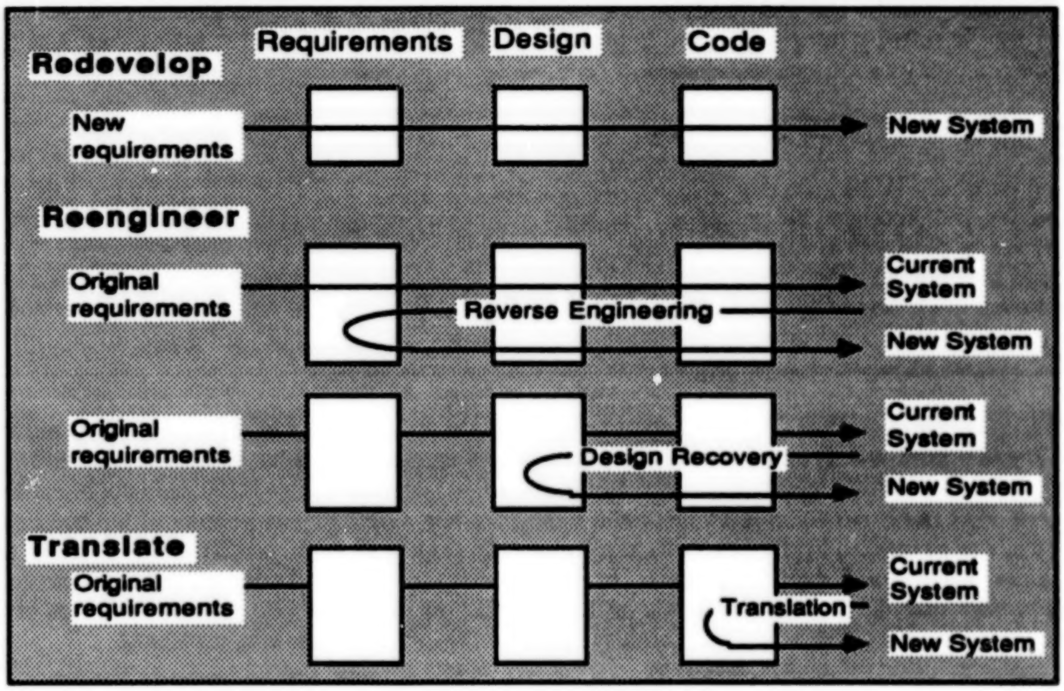

Requirements
Design
Code
Redevelop
New requirements
New System
Reengineer
Original requirements
Reverse Engineering
Current System
New System
Original requirements
Design Recovery
Current System
New System
Translate
Original requirements
Translation
Current System
New System

Figure 1. Alternative Solutions

- Poor existing control flow is translated into poor control flow.

- Poor existing data structures remain poor data structures.

- Input/output translation usually produces hard to read "unnatural"code in the new language.

- Translation does not take advantage of the code and data packaging techniques available in the newer languages. Attempts to automatically translate some FORTRAN programs to Ada have failed.

Reengineering is the combination of "reverse engineering" a working software system and then "forward engineering" a new system based on the results of the reverse engineering. Forward engineering is the standard process of generating software from "scratch." It is composed of the life cycle phases such as requirements, architectural design, detailed design, code development, testing, etc. In each phase, certain products are required and the activities which produce them are defined. Each product is required to be complete and consistent. To progress forward to a new phase normally requires a new representation of the products which involve more detail such as new derived requirements, design decisions, trade off evaluation between alternative approaches, etc. Finally, code is developed which is the most complete, consistent, and detailed representation of the required product.

Reverse engineering is the reverse of forward engineering. It is the process of starting with existing code and going backward through the software development life cycle. Life cycle products are, therefore, obtained by abstracting from more detailed representations to more abstract ones. This process should proceed much faster than forward engineering since all of the details required are available. Reverse engineering starts with the most detailed representation, which has also proven to be complete and consistent since it can currently do the job required. Developing products in reverse involves abstracting out only the essential information and hiding the non-essential details at each reverse step.

How far to go backward in the reverse engineering process before it is stopped and forward engineering begins is a critical question and involves trade offs. It is important to understand all of what the program does, all of the information it handles, and the control flow since these are probably required to get the job done. This implies taking the reverse process far enough to understand *what* the "as is" program is. This is usually more significant than *how* the program does its job since the *how* is usually the part that will be changed in any following forward engineering process.

What a program does is called its *requirements*. *How* it meets those requirements is its *design*. For a reverse engineered program it is the design that will be updated more often than *what* the program will do. Modern software engineering techniques and technologies such as user interfaces, database management, memory utilization, data structuring, packages, objects, etc. will affect the design, not *what* the program does. Therefore, once it is understood what the program does and what is obsolete, then the forward engineering process can begin with confidence.

Reverse engineering is referred to as "design recovery" when the reverse engineering process stops at the recovery of the design of the implementation, rather than proceeding on to a higher level of abstraction to include the recovery of the requirements. The basic process of this level of design recovery involves recovery of information about the code modules and the data structures in an existing program. This information will support the programmer/analyst who is maintaining an unfamiliar large FORTRAN program, upgrading it for maintainability, or converting it to another target language.

However, a better job of redesigning a program can be accomplished with requirements recovery than with design recovery. To carry the reverse engineering process beyond design recovery to requirements recovery is difficult and requires higher levels of domain knowledge to do the abstractions. The *whys* of the requirements, design, and implementation can only be provided by someone very familiar with the program and the domain. This level of expertise is often very difficult to find and have dedicated to the reengineering process. For this reason, the methods and tools that the STB has developed initially assume reverse engineering only to the design recovery stage. Future development will be based on feedback from the JSC software engineering community. The current standards, methods, tools, and environment are all designed to be sufficiently flexible and extendible to enable the strategies to be extended to cover the full spectrum of reverse engineering.

The overriding philosophy of this planned reverse engineering process is to capture the total software implementation in an electronic form. This includes source code, documentation, databases, etc. Figure 2 illustrates the progression of data structures from COMGEN-compatible code (see section 4.0) to reengineered code. This progression in electronic form ensures that the total consistent and complete requirements representation is available. Software tools are provided to support the generation of the more abstract products required for engineering in reverse as well as capturing rationale and decisions of the engineer. By the continuing process of abstracting the information about the program into the different representations, the engineer can remain more confident that information is not being lost or inadvertently "falling through the cracks."

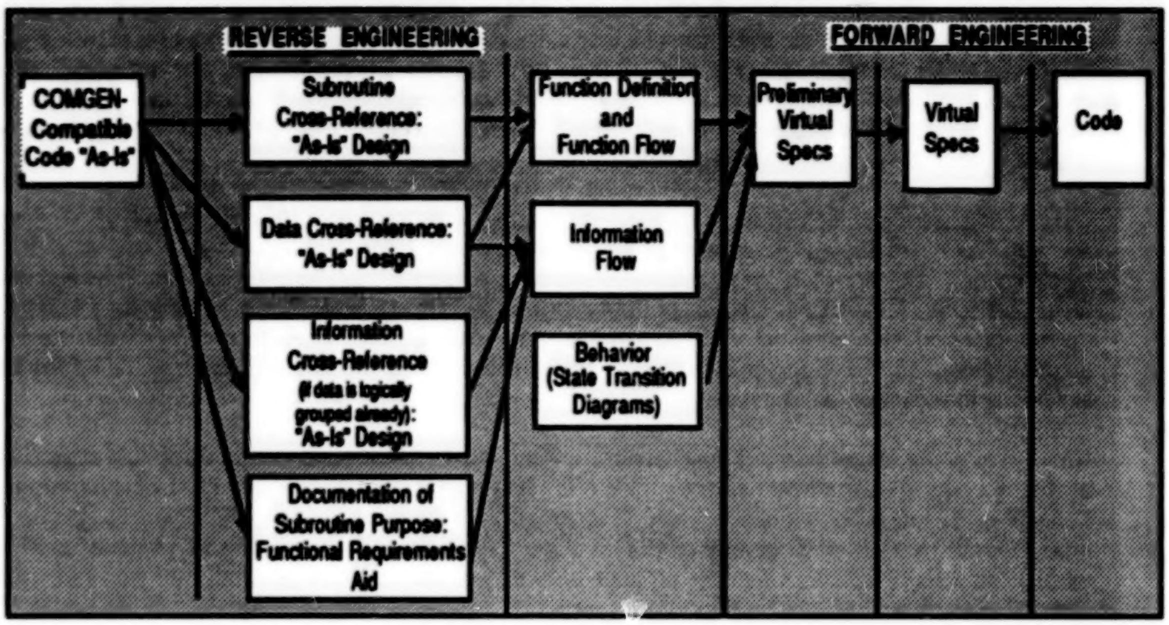

REVERSE ENGINEERING
FORWARD ENGINEERING
COMGEN-Compatible Code "As-Is"
Subroutine Cross-Reference: "As-Is" Design
Data Cross-Reference: "As-Is" Design
Information Cross-Reference (if data is logically grouped already): "As-Is" Design
Documentation of Subroutine Purpose: Functional Requirements Aid
Function Definition and Function Flow
Information Flow
Behavior (State Transition Diagrams)
Preliminary Virtual Specs
Virtual Specs
Code

Figure 2. Data Structure Progression

SOFTWARE TECHNOLOGY BRANCH'S REENGINEERING HISTORY

In the early 1970's, the Mission Planning and Analysis Division's (MPAD) Software Development Branch and TRW/Houston developed a tool, called COMGEN, that began as a COMMON block specification statement generator. It grew to include many other functions as new techniques were developed. Later COMGEN was broken up into a continually evolving set of tools with common data interface structures. This tool set supports the maintenance of FORTRAN programs today on Unisys and multiple Unix systems. People still refer to this tool set as COMGEN tools, and a program that complies with the MPAD standard COMMON concept as a COMGEN-compatible program. [1,2,3]

In the 1970's, MPAD performed a lot of software reengineering to meet the goal of combining many of the independently developed engineering programs, each with its own input/output formats. Many of the modern concepts such as separation of input/output processing from the applications, databases, data structures, packages, generics, objects, etc. were recognized and simulated to some degree. They were not called by the modern names, of course, but the design engineers were trying to do good engineering, modularization, and data handling. Even though these techniques were known in the 1970's, they are just now really becoming popular because of newer technologies such as database management systems, user interface tools sets, and modern languages that actually embed and enforce good software engineering practices.

In the late 1980's, some of the personnel and the functions of the Software Development Branch were reorganized into the newly created STB, the Software Technology Branch. The STB's reengineering history has put JSC in a better position with respect to the maintainability of its older software than many other organizations. The positive results of this experience include the following:

- Most of the software is reasonably modular.
- The data has some structure.
- Most of the software at JSC is reasonably compatible with the STB's tools, including the in-line documentation.
- The large complex programs that support many simulations have considerable software reuse and information sharing.

MAINTENANCE STRATEGIES

The strategies presented in this document are intended to help with design recovery in support of programmer/analysts who are required to maintain large FORTRAN programs that they did not develop. In addition, these strategies are intended to support reengineering of existing FORTRAN code into modern software engineering structures, which are then easier to maintain and which allow a fairly straight forward translation into other target languages. The STB is proposing standards, methods, and an integrated software environment based upon the significant set of tools built to develop and maintain FORTRAN code for the Space Shuttle. [4,5,6,7,8] The environment will support these structures and practices even in areas where the language definition and compilers do not enforce good software engineering practices.

New FORTRAN Standards

New standards, which allow modern software engineering constructs to be used in FORTRAN 77, have been defined by the STB. [5] These standards are added to existing standards defined by the former MPAD and still in use in the mission planning and analysis domain. The goal of the new standards is to improve maintainability and permit relatively automated translations to newer languages. In table 1, the standards and their benefits are summarized. These standards address documentation, longer variable names, modern control flow structures, grouping subprograms together as virtual packages, data structuring, and input/output encapsulation in separate subprograms. Where FORTRAN 77 does not provide the constructs, virtual constructs are provided along with a tool environment to support their development and maintenance. The existing core of FORTRAN programmers should have little problem with the standards and new FORTRAN code should adhere to them from the start.

Table 1. Standards Summary

Standard	Benefit
Documentation Header statement before code blocks Requirements in CD1 statements Rationale in CD7 statements Virtual package identification	 Understandability Understandability and traceability Design knowledge capture Maintenance
Longer, more meaningful variable names	Understandability
Modern control flow structures Block DO DO WHILE	Maintenance and understandability
Grouping subprograms into virtual packages	Higher level of abstraction, understandability
Data structuring Preferred use of calling parameters Controlled use of COMMON blocks	 Maintenance Maintenance INCLUDE COMMON database concept
Preferably encapsulate input/output in separate subprograms	Maintenance and support to future conversions

Design Recovery and Reengineering Methodology

The reengineering methodology defines the steps, the skills required, and guidelines on how far to reverse engineer before deciding to rebuild. The key goal is to update to modern technology and software engineering concepts without losing required functions and data. Methods are provided that have the flexibility to meet multiple levels of conversion, each of which improves maintainability. Figure 3 illustrates five methods. [6] Method 1 converts an arbitrary FORTRAN program to COMGEN-compatible FORTRAN, which provides in-line documentation, data structure, and unique data names within a COMMON structure. Method 2 converts software already in this format to the new "standard" FORTRAN with a more Ada-like structure that is ready for a mostly automated translation by Method 3 to a target language that embeds software engineering principles. Alternatively, COMGEN-compatible programs can be converted directly to a target language like Ada by Method 4. Although it is easier to convert a FORTRAN program when the code already meets the standard COMMON concept, commonly known as COMGEN-compatible, arbitrary FORTRAN can be directly converted to a target language by Method 5.

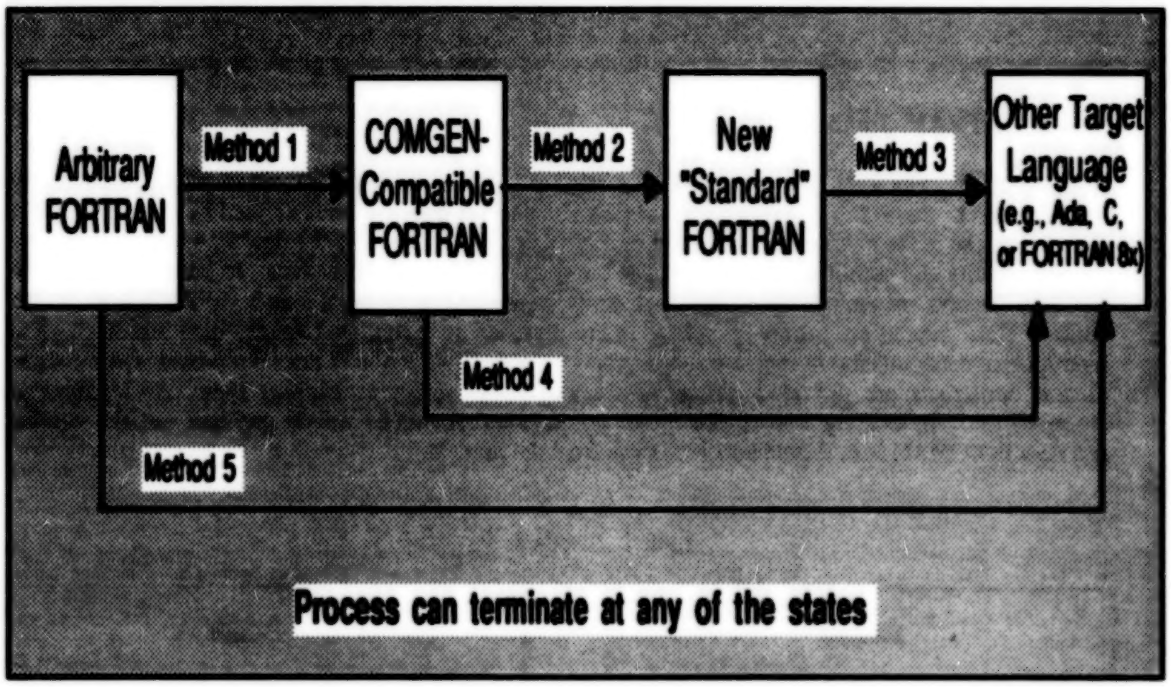

Arbitrary FORTRAN
Method 1
COMGEN-Compatible FORTRAN
Method 2
New "Standard" FORTRAN
Method 3
Other Target Language (e.g., Ada, C, or FORTRAN 8x)
Method 4
Method 5
Process can terminate at any of the states

Figure 3. Reengineering Methods

Environment to Support Design Recovery and Reengineering

The STB's reengineering environment [7] will be built around three components: standards, methods, and tools that support the standards and the methods. It will contain modified versions of the tools used to support the current JSC FORTRAN programs plus commercial off-the-shelf (COTS) tools and additional custom-built tools. The intent is to get an environment out into use in JSC's maintenance community to provide support for upgrading FORTRAN programs in terms of maintainability in the near-term, then to extend the functionality of the tool set and environment in response to feedback from the programmers/ analysts. Later versions of the environment may have extensions to handle subject programs written in C, Ada, or even HAL/S, according to requests from the user community.

The environment will be designed with stable interfaces defined to provide for the maximum degree of seamlessness that is desirable. It is doubtful that COTS tools can be integrated seamlessly into the environment as no standard interfaces have yet been established for either user interface or data interface (as opposed to data exchange). The tools will be integrated at the front end by a user interface and behind the screen by two logical databases, one containing data passed to and from the tools and the other containing the original and modified source code as shown in figure 4.

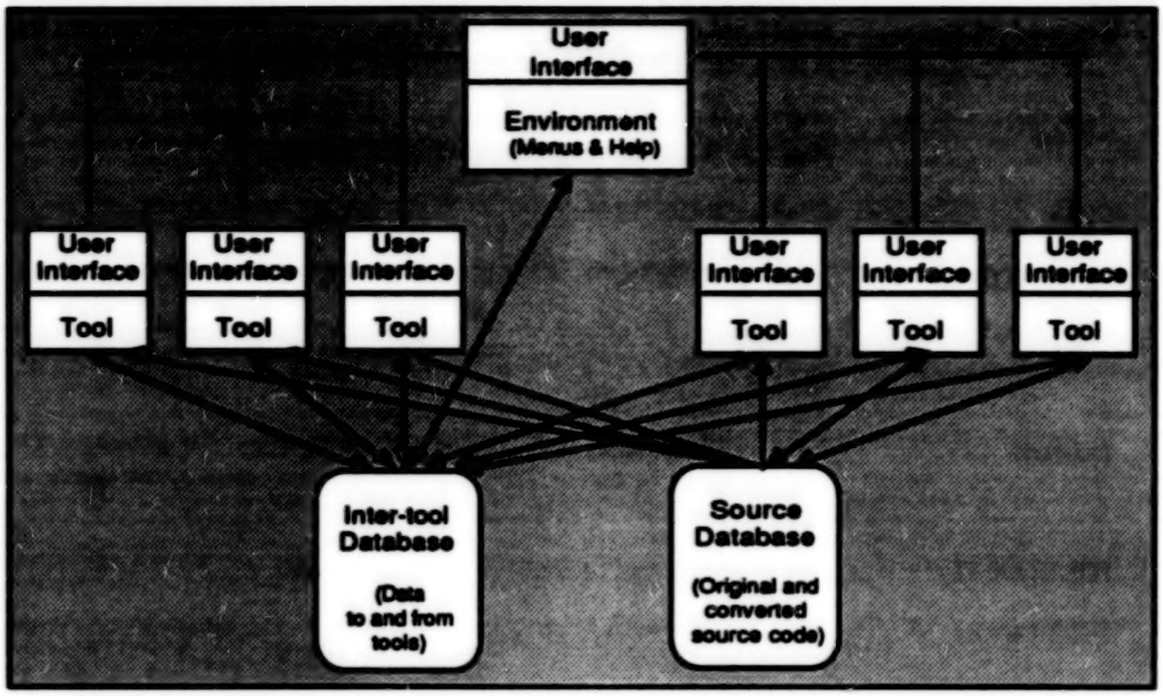


Figure 4. Conceptual Architecture of the Design Recovery and Reengineering Environment

The environment will not be a completely automated environment since much work will still have to be done by a programmer/analyst. A person must be in the loop to provide the required puzzle-solving skills that are beyond the capabilities of state-of-the-practice tools. However, as an experience base is accrued in design recovery and reengineering, knowledge-based capabilities can be added to the environment.

CONCLUSIONS

JSC has a large amount of exsting code in FORTRAN that embodies domain knowledge and required functionality. This code must be maintained and eventually translated to more modern languages. Three primary alternative solutions have been identified to address the maintenance problems of these old FORTRAN programs: complete redevelopment of the programs, code translation to a more modern language or version of a language, and reengineering. Complete redevelopment is effective but very costly. Simple code translation is cheap, but usually ineffective since seldom do the old systems incorporate modern software engineering concepts such as good data structuring, good control structuring, packages, objects, etc., that should be present in the new system. Modern languages such as Ada have constructs for representing these features, but translators cannot determine these features in the original code to map them into the new system. Reengineering is being recognized as a viable option because the old systems, in spite of obsolete technology, do contain all of the required functionality and can get the job done. However, at the present time there are only a few expensive Computer Aided Software Engineering (CASE) tools and no total system environment available in the COTS market to support reengineering FORTRAN programs.

The STB maintenance strategies provide standards, methods, and a tool environment for upgrading current FORTRAN systems without losing the embedded engineering knowledge and at a lower cost than for complete redevelopment of the program. A useful environment for reengineering FORTRAN software can be built fairly quickly by building upon the existing FORTRAN development and maintenance tools, COTS products, new software and hardware technologies, plus current research into reuse, design recovery, and reengineering. This environment will support reengineering existing FORTRAN code into more maintainable forms that can also be readily translated into a modern language including newer versions of FORTRAN.

GLOSSARY

arbitrary FORTRAN	FORTRAN program that is not compatible with the COMGEN standards long in place for JSC's mission planning and analysis domain.
COMGEN-compatible	FORTRAN program that is compatible with the COMGEN standards long in place for JSC's mission planning and analysis domain. [1]
COTS	Commercial-Off-The-Shelf
design recovery	Reverse engineering, the first step for maintenance or reengineering.
environment	Instantiation of a framework, i.e., an integrated collection of tools. It may support one or more methodologies and may also provide a framework for third party tools.
framework	Software system to integrate both the data and the control of new and existing tools; usual components include a user interface, object management system, and a tool set.
FORTRAN 77	ANSI standards for FORTRAN in effect in June 1990.
FORTRAN 8x	Future ANSI standards for FORTRAN; expected to be approved and released soon; draft standards have been circulated; unofficially called FORTRAN 90.
forward engineering	Process of developing software from "scratch," through the phases of requirements, design, and coding.
package	"A collection of logically related entities or computational resources" (Booch[9]).
reengineering	"The examination and alteration of a subject system to reconstitute it in a new form and the subsequent implementation of the new form" (Chikofsky and Cross [10]); combination of reverse engineering and forward engineering.
reverse engineering	"The process of analyzing a subject system to identify the system's components and their interrelationships and create representations of the system in another form or at a higher level of abstraction" (Chikofsky and Cross [10]); the first step of maintenance or reengineering; reverse of forward engineering; process of starting with existing code and going backward through the software development life cycle.
software maintenance	Process of modifying existing operational software while leaving its primary functions intact (Boehm [11]).
subject program	Program that is being maintained or reengineered.
virtual package	Package concept as defined by Booch [9], but implemented either in Ada, which enforces the concept, or in a language in which the concept must be supported procedurally.

REFERENCES

[1] Braley, Dennis: *Computer Program Development and Maintenance Techniques*. NASA IN 80-FM-55, NASA Johnson Space Center (Houston, TX), November 1980.

[2] Braley, Dennis: *Automated Software Documentation Techniques*. NASA Johnson Space Center (Houston, TX), April 1986.

[3] Braley, Dennis: *Software Development and Maintenance Aids Catalog*. NASA IN 86-FM-27, NASA Johnson Space Center (Houston, TX), October 1986.

[4] Fridge III, Ernest: *Maintenance Strategies for Design Recovery and Reengineering: Executive Summary and Problem Statement*. Volume 1. NASA Johnson Space Center (Houston, TX), June 1990.

[5] Braley, Dennis: *Maintenance Strategies for Design Recovery and Reengineering: FORTRAN Standards*. Volume 2. NASA Johnson Space Center (Houston, TX), June 1990.

[6] Braley, Dennis; and Plumb, Allan: *Maintenance Strategies for Design Recovery and Reengineering: Methods*. Volume 3. NASA Johnson Space Center (Houston, TX), June 1990.

[7] Braley, Dennis; and Plumb, Allan: *Maintenance Strategies for Design Recovery and Reengineering: Concepts for an Environment*. Volume 4. NASA Johnson Space Center (Houston, TX), June 1990.

[8] George, Vivian; and Plumb, Allan: *A Method for Conversion of FORTRAN Programs*. Barrios Technology, Inc. (Houston, TX), March 1990.

[9] Booch, G.: *Software Engineering with Ada*. Benjamin/Cummings Publishing Co., Inc. (Menlo Park, CA), 1983.

[10] Chikofsky, E. J.; and Cross II, J. H.: "Reverse Engineering and Design Recovery: A Taxonomy." *IEEE Software*, January 1990.

[11] Boehm, B. W.: *Software Engineering Economics*. Prentice-Hall (Englewood Cliffs, NJ), 1981.

N91

24048

UNCLAS

N91-24048

TRANSPORTABLE APPLICATIONS ENVIRONMENT (TAE) PLUS: A NASA USER INTERFACE DEVELOPMENT AND MANAGEMENT SYSTEM

Martha R. Szczur

NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
mszczur@postman.gsfc.nasa.gov

ABSTRACT

The Transportable Applications Environment Plus (TAE™ [illegible], developed at NASA's Goddard Space Flight Center, is a portable *What You See Is What You Get* (WYSIWYG) user interface development and management system. Its primary objective is to provide an integrated software environment that allows interactive prototyping and development of graphical user interfaces, as well as management of the user interface within the operational domain. TAE Plus is being applied to many types of applications, and this paper discusses what TAE Plus provides, how the implementation has utilized state-of-the-art technologies within graphic workstations, and how it has been used both within and outside NASA.

BACKGROUND

Emergence of graphical user interfaces

With the recent emergence of sophisticated graphic workstations and the subsequent demands for highly interactive systems, designing and developing good user interfaces has become more complex and difficult. Prior to the graphic workstations, the application developer was primarily concerned with developing user interfaces for a single monochrome 80x24 alphanumeric character screen with keyboard user entry. With high resolution bit-mapped workstations, the user interface designer has to be cognizant of multiple window displays, the use of color, graphical objects and icons, and various user selection techniques (e.g., mouse, trackball, tablets).

High resolution graphic workstations also provide system developers with the opportunity to rethink and redesign the user interfaces (UI) of their next generation applications. For instance, in a command and control environment, many processes run simultaneously to monitor a particular operation. With modern graphic workstations, time critical information concerning multiple events can be displayed concurrently on the same screen, organized into different windows in a variety of graphical and textual presentations. As today's workstations inspire more elaborate user interfaces, the applications which utilize their graphics capabilities increase in complexity. Prototyping different user interface designs, thus, becomes an increasingly important method for developing and communicating concepts and requirements for an application.

Role of prototyping at GSFC

Prototypes can be constructed with various levels of sophistication and fidelity. At their simplest they are visual mockups of the user interface. A prototype can also be a dynamic sequence of events, with simulated control between steps. They can even be a working model of a system, which can evolve into an operational system or be used for research purposes.

Within the government environment, prototypes also can play the important role of communicating specifications from government agency to the contractor, as well as to validate contractor interpretation and design approaches by reviews from the targeted government user. Prototyping key concepts and salient features of proposed user interface standards, applied in typical operations scenarios, greatly enhances the users' ability to respond and have their concerns understood. Thus, including prototyping as a step in the application development cycle can ensure user acceptance of the final operational application.

Requirements for a prototyping-to-operational development environment

To support our development methodology we wanted to establish an integrated environment that allows prototyped user interfaces to evolve into operational applications. This environment would satisfy the following objectives:

- separate the user interface from the application,
- provide tools to allow interactive design/change/save of user interface elements,
- take advantage of the latest hardware technology,
- support rapid prototyping,
- manage the user interface,
- allow integrated management of multiple, asynchronously-active processes,
- develop tools for increasing application development productivity,
- provide the application with runtime services, and
- allow portability to different computing environments.

Building on existing technology

Many of these objectives were addressed in the early 1980's when GSFC recognized that most large-scale space applications, regardless of function, required software to support human-computer interactions and application management. This lead to the design and implementation of the Transportable Applications Executive (now, referred to as TAE *Classic*), which abstracts a common core of system service routines and user dialog techniques used by all applications [Ref. 1]. Over the years, TAE Classic has matured into a powerful tool for quickly and easily building and managing consistent, portable user inter-faces, but only for the standard alphanumeric terminal. When the requirement to support graphical user interfaces emerged, TAE Classic was examined as a potential *building block*. It was determined that it had a sufficiently flexible architecture and data structure to accommodate the extensions that would be needed to support user interface development within the graphic workstation environment.

WHAT DOES TAE PLUS PROVIDE?

To meet the defined goals, services and tools were developed for creating and managing window-oriented user interfaces. It became apparent, due to the flexibility and complexity of graphical user interfaces, that the design of the user interface should be considered a separate activity from the application program design. The interface designer can then incorporate human factors and graphic art techniques into the user interface design. The application programmer needs only to be concerned about what results are returned by the user interaction and not the look of the user interface.

In support of the user interface designer, an interactive *WorkBench* application was implemented for manipulating interaction objects ranging from simple buttons to complex multi-object panels. As illustrated in Figure 1, after designing the screen display, the WorkBench saves the specification of the user interface in resource files, which can then be accessed by application programmers through a set of runtime services, Window Programming Tools (WPTs). Guided by the information in the resource files, the routines handle all user interactions. The WPTs utilize the standard MIT X Window System™ to communicate with the graphic workstations. [Ref. 2] As a further aid to the UI developer, the WorkBench provides an option to generate the source code which will display and manage the designed user interface. This gives the programmer a working template into which application-specific code can be added.

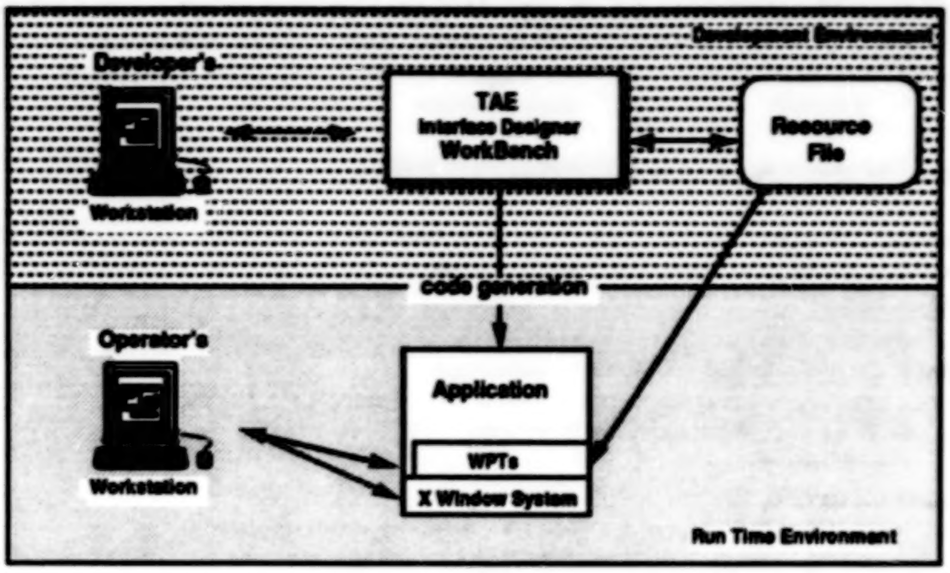
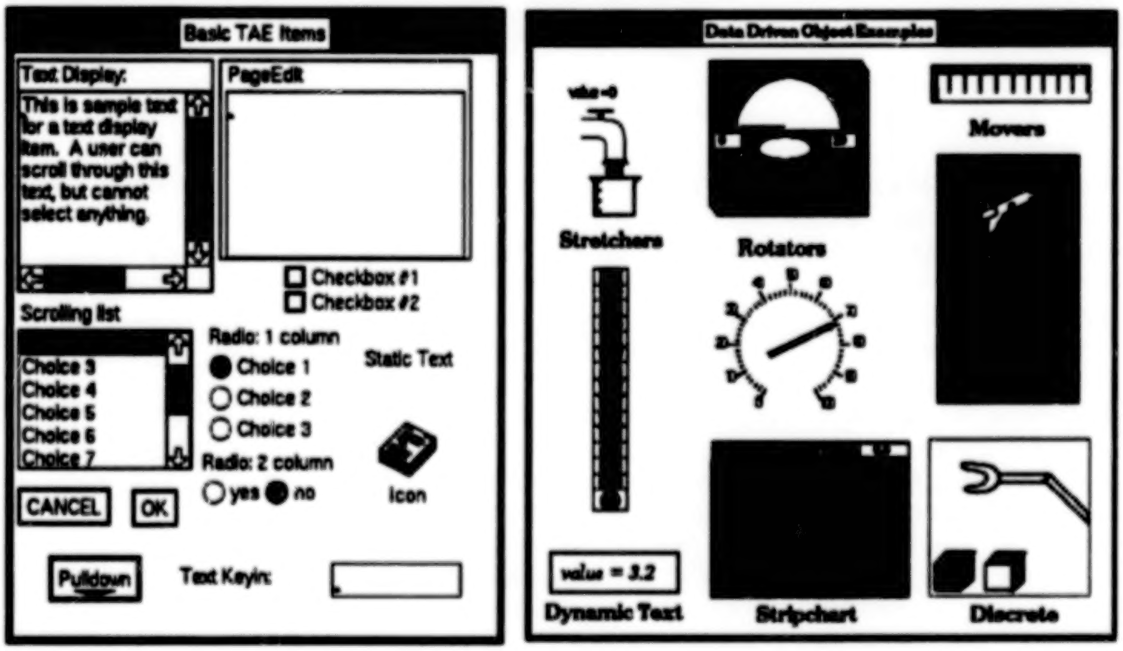

Development Environment
Developer's
Workstation
TAE
Interface Designer
WorkBench
Resource
File
code generation
Operator's
Workstation
Application
WPTs
X Window System
Run Time Environment

Figure 1. TAE Structure

INTERACTION OBJECTS AS BUILDING BLOCKS

The basic building blocks for developing an application's graphical user interface are a set of interaction objects. All visually distinct elements of a display that are created and managed using TAE Plus are considered to be interaction objects and they fall into three categories: user-entry objects, information objects, and data-driven objects. *User-entry objects* are mechanisms by which an application can acquire information and directives from the end user. They include radio buttons, check boxes, text entry fields, scrolling text lists, pulldown menus and push buttons. *Information objects* are used by an application to instruct or notify the user, such as contextual on-line help information displayed in a scrollable static text object or brief status error messages displayed in a bother box. *Data-driven objects* are vector-drawn graphic objects


Basic TAE Items
Text Display:
PageEdit
☐ Checkbox #1
☐ Checkbox #2
Scrolling list
Choice 3
Choice 4
Choice 5
Choice 6
Choice 7
Radio: 1 column
● Choice 1
○ Choice 2
○ Choice 3
Static Text
Radio: 2 column
○ yes ● no
Icon
CANCEL
OK
Pulldown
Text Keyin:
Data Driven Object Examples
value=0
Stretchers
Rotators
Movers
value = 3.2
Dynamic Text
Stripchart
Discrete

This is sample text for a text display item. A user can scroll through this text, but cannot select anything.

Figure 2. TAE Plus User Interface Interaction Objects

which are linked to an application data variable; elements of their view change as the data values change. Examples are dials, thermometers, and strip charts. When creating user dialogues, these objects are grouped and arranged within *panels* (i.e., windows) in the WorkBench.

The use of interaction objects offers the application designer/programmer a number of benefits with the expected payoff of an increase in programmer productivity. The interaction objects provide a consistent look and feel for the application's user interface, which translates into reduced end-user training time, more attractive screens, and an application which is easier to use. Another key benefit is that since the interaction objects have been thoroughly tested and debugged, the programmer is able to spend more time testing the application and less time verifying that the user interface behaves correctly. This is particularly important considering the complexity of some of the objects, and the programming effort it would take to code them from scratch. Refer to Figure 2 for a sample of the TAE Plus interaction objects.

TAE PLUS WORKBENCH

The WorkBench provides an intuitive environment for defining, testing, and communicating the look and feel of an application system. Functionally, the WorkBench allows an application designer to dynamically lay out an application screen, defining its static and dynamic areas. The tool provides the designer with a choice of pre-designed interaction objects and allows for tailoring, combining and rearranging of the objects. To begin the session, the designer needs to create the base panel (i.e., window) into which interaction objects will be specified. The designer specifies presentation information, such as the title, font, color, and optional on-line help for the panel being created. The designer defines both the presentation information and the context information of all interaction items to reside in the panel by using the item specification window (refer to Figure 3). For icon support, the WorkBench has an icon

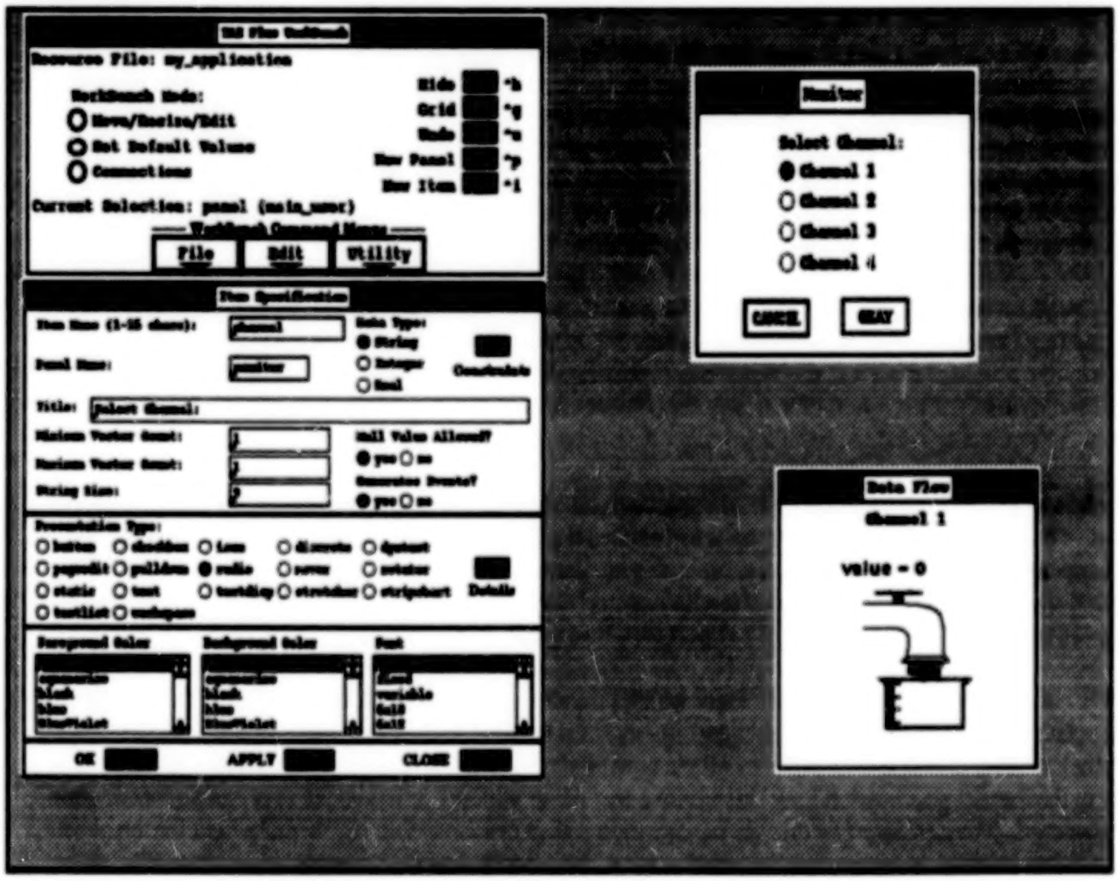


Figure 3. Building a user interface with the WorkBench

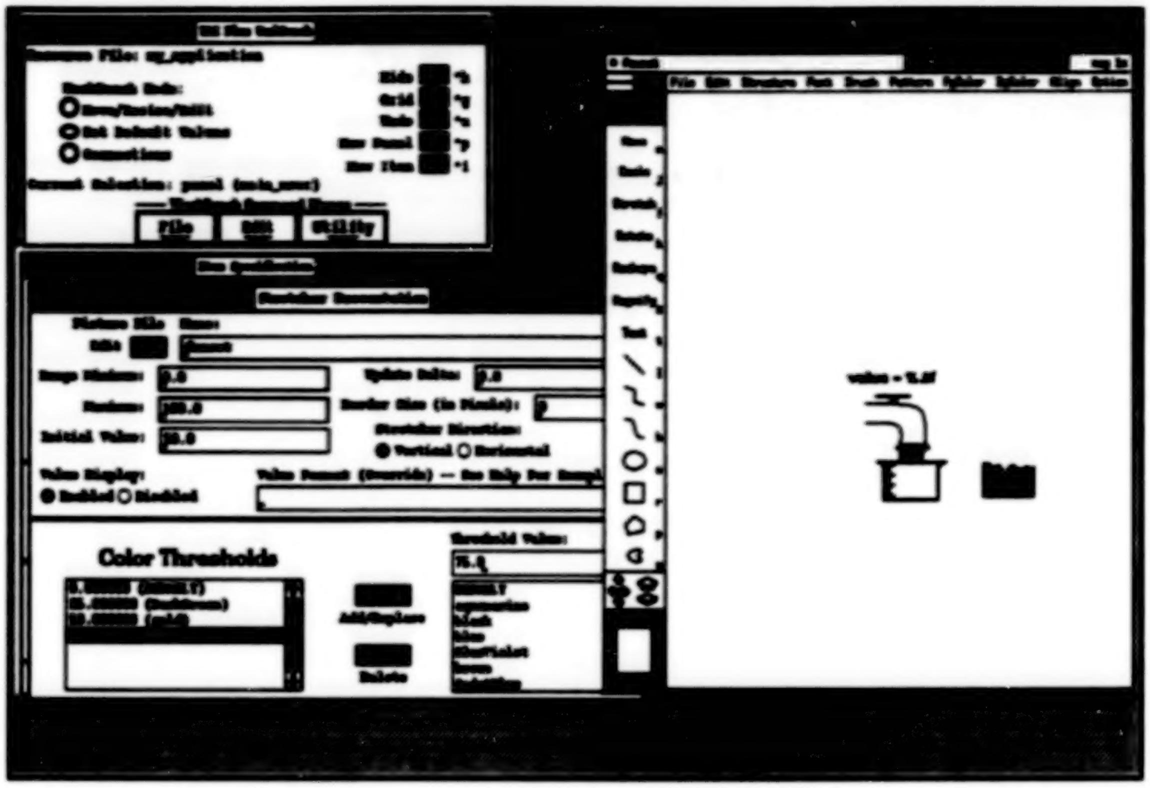


Figure 4. Creating a stretcher data-driven object

editor, within which an icon can be drawn, edited and saved. As the UI designer moves, resizes, and alters any of the item's attributes, the changes are dynamically reflected on the display screen.

The designer also has the option of retrieving *palettes* of previously created items. The ability to reuse interaction objects saves programming time, facilitates experimenting with different combinations of items in the prototyping process, and contributes to standardization of the application's look and feel. If an application system manager wanted to ensure consistency and uniformity across an entire application's UI, all developers could be instructed to use only items from the application's palette of common items.

When creating a data-driven object, the designer goes through a similar process by setting the associated attributes (e.g., color thresholds, maximum, minimum, update delta) in the specification panels. To create the associated graphics drawing, the WorkBench provides a drawing tool within which the static background and dynamic foreground of a data-driven object can be drawn, edited, and saved. Figure 4 shows the drawing tool being used to create a *stretcher* data-driven object.

Most often an application's UI will be made up of a number of related panels, sequenced in a meaningful fashion. Through the WorkBench, the designer defines the interface *connections*. These links determine what happens when the user selects a button or a menu entry. The designer attaches *events* to interaction items and thereby designates what panel appears and/or what program executes when an event is triggered. Events are triggered by user-controlled I/O peripherals (e.g., point and click devices or keyboard input).

TAE Plus also offers an optional help feature which provides a consistent mechanism for supplying application-specific information about a panel and any interaction items within the panel. In a typical session, the designer elects to edit a help file after all the panel items have been designed. Clicking on the edit help option in the Panel Specification Panel brings up a text editor window in which the appropriate information can be entered. The designer can then define any button item or icon item to be the help item for the panel (in this scenario it would be the help icon in the panel "Monitor"). During the application operation, when

ORIGINAL PAGE IS
OF POOR QUALITY

the end-user clicks on the question mark item, the cursor changes to a question mark symbol (?). The end-user then clicks on the panel itself or any item in the panel to bring up a help panel containing the associated help text.

Having designed the layout of panels and their attendant items and having threaded the panel and items according to their interaction scenario, the designer is able to preview (i.e., rehearse) the interface's operation from the WorkBench. With this potential to test drive an interface, to make changes, and to test again, iterative design becomes part of the prototyping process. With the rehearsal feature, the designer can evaluate and refine both the functionality and the aesthetics of a proposed interface. After the rehearsal, control is returned to wherever the designer left off in the WorkBench and the designer can either continue with the design process or save the defined UI in a resource file.

Developing software with sophisticated user interfaces is a complex process, mandating the support of varied talents, including human factors experts and application program specialists. Once the UI designer (who may have limited experience with actual code development) has finished the UI, he/she can turn the saved UI resource file over to an experienced programmer. As a further aid to the application programmer, the WorkBench has a "generate" feature, which produces a fully annotated and operational body of code which will display and manage the entire WorkBench-designed UI. Currently, source code generation of C, Ada, FORTRAN and TCL are supported, with bindings for C++ expected in a future release of TAE Plus. The programmer can now add additional code to this template and make a fully functional application. Providing these code stubs helps in establishing uniform programming method and style across large applications or within a family of interrelated software applications.

WINDOW PROGRAMMING TOOLS (WPTs)

The Window Programming Tools (WPTs) are a package of application program callable subroutines used to control an application's user interface. Using these routines, applications can define, display, receive information from, update and/or delete TAE Plus panels and interaction objects. WPTs support a modeless user interface, meaning a user can interact with one of a number of interaction objects within any one of a number of displayed panels. In contrast to sequential mode-oriented programming, modeless programming accepts, at any instance, a number of user inputs, or *events*. Because these multiple events must be handled by the application program, event-driven programming can be more complex than traditional programming. The WorkBench's auto-generation of the WPT event loop reduces the risk of programmer error within the UI portion of an application's implementation.

Wpt_AddEvent	Add other sources for input/output/exception
Wpt_BeginWait	Display busy indicator cursor
Wpt_CloseItems	Close Items on a Panel
Wpt_ConvertName	Get the X Id of a named window
Wpt_Endwait	Stop displaying busy indicator cursor
Wpt_Init	Initializes interface to X Window System
Wpt_ItemWindow	Gets the window id of the window containing a parameter
Wpt_MissingVal	Indicates if any values are missing
Wpt_New Panel	Displays a user interface panel
Wpt_NextEvent	Gets next panel-related event
Wpt_PanelErase	Erases the displayed panel from the screen
Wpt_PanelMessage	Displays message in "Bother Box"
Wpt_PanelReset	Resets object values to initial values
Wpt_PanelTopWindow	Gets panel's parent shell window id
Wpt_PanelWidgetId	Return the Widget Id of a Wpt Panel Widget
Wpt_PanelWindow	Returns the X Id of a panel
Wpt_ParmReject	Generates a rejection message for a given value
Wpt_ParmUpdate	Updates the displayed values of an object
Wpt_Pending	Check if a WptEvent is pending from X, Parm or file.
Wpt_RemoveEvent	Remove a previously registered event
Wpt_SetTimeOut	Set/Cancel timeout for gathering Wpt events.
Wpt_ViewUpdate	Updates the view of a parameter on a displayed panel

Figure 5. The Window Programming Tools (WPTs)

As mentioned earlier, the WPT package utilizes the MIT X Window System as its base windowing system. One of the strengths of X is the concept of providing a low-level abstraction of windowing support (Xlib), which becomes the base standard, and a high-level abstraction (X toolkits), which has a set of interaction objects (called "widgets" in the X world) that define elements of a UI's look and feel. The current version of TAE Plus (V4.1) operates with the X11R3 and X11R4 version using the X Toolkit and HP widget set delivered with the X software. Due to the growing acceptance of the Open Software Foundation's Motif ™ user interface style as a defacto industry standard, the next release of TAE Plus (V5.0) will be based on the Motif software.

The WPTs also provide a buffer between the application program and the X Window System services. For instance, to display a WorkBench-designed panel, an application makes a single call to Wpt_NewPanel (using the *panel name* specified in the WorkBench). This single call translates into a function that can make as many as 50 calls to X Window System routines. For the majority of applications, the WPT services and objects supported by the WorkBench provide the necessary user interface tools and save the programmer from having to learn the complexities of programming directly with X. This can be a significant advantage, especially when considering the learning curve differential between 26 WPT routines versus over 400 X Toolkit intrinsics and over 200 Xlib services. Refer to Figure 5 for a sample list of the WPTs.

IMPLEMENTATION

The TAE Plus architecture is based on a separation of the user interaction management from the application-specific software. The current implementation is a result of having gone through several prototyped and beta versions of a WorkBench and user interface support services during the 1986-89 period, as well as building on the TAE Classic structure.

The "Classic" portion of the TAE Plus code (≈ 60,000 LOC) is implemented in the C programming language. In selecting a language for the WorkBench and the WPT runtime services, we felt a "true" object-oriented language would provide us with the optimum environment for implementing the TAE Plus graphical user interface capabilities. (See Chapter 9 of Cox [Ref. 4] for a discussion on the suitability of object-oriented languages for graphical user interfaces.) We selected C++ [Ref. 5] as our implementation language for several reasons [Ref. 6]. For one, C++ is becoming increasingly popular within the object-oriented programming community. Another strong argument for using C++ was the availability of existing, public domain, X-based object class libraries. Utilizing an existing object library is not only a cost saver, but also serves as a learning tool, both for object-oriented programming and for C++. Delivered with the X Window System is the *InterViews* C++ class library and a drawing utility, *idraw*, both of which were developed at Stanford University. [Ref. 7] The *idraw* utility is a sophisticated direct manipulation C++ application, which we integrated into the WorkBench to support creating, editing and saving the graphical data-driven interaction objects.

PORTABILITY ISSUES

Throughout the design and development of TAE Plus, one of our primary goals has been to be "portable" over a wide range of hardware platforms. It is a requirement that TAE Plus operate on various UNIX systems and VAX/VMS. There are three primary software areas identified to be the most nonportable -- file manipulation, process control and interprocess communication. The software modules to support these areas are localized and tailored to the individual operating system of each host environment. With the proper use of tailored *include* files, TAE Plus ports between workstations with few problems.

When porting among different hardware platforms, the host system's method of storing binary information is always of key concern. Since TAE Plus's resource files are binary, we provide a utility to produce a straight ASCII equivalent of the file. This file can then be transferred to any platform that accepts the ASCII character set and then converted back into a binary file, which can be read by TAE Plus applications (including the WorkBench) operating on the target platform.

As mentioned earlier in this paper, the C language was selected for implementing TAE Classic and it has proven to be an efficient and standard language across different hardware platforms. The C++ code has proven to be less portable than anticipated. There are several differences, even syntactical, among the various C++ compilers. Therefore it was decided to initially limit our support to just two compilers. They are the GNU C++ and the OASYS Designer C++ compiler. With the recent release of the AT&T V2.0 C++ compiler, C++ is becoming more standardized and the compiler issue is expected to dissipate as vendor's offer C++ as one of their standard language compilers.

The single most important factor contributing to the portability of TAE Plus is the X Window System. Generally, if a graphic workstation supports the Xlib and X Toolkit and operates either UNIX or VMS, TAE Plus can be ported to it with reasonable ease.

AVAILABILITY AND USER SUPPORT

After two years of prototyping and developing beta versions of the TAE Plus, an *industrial strength* version of TAE Plus (Version 4.1) was released in February 1990. It is available for public distribution, at a minimal license fee, from the Center of Software Management and Information Center (COSMIC), a NASA distribution center. While TAE Plus base development and testing is done on a Sun workstation under UNIX within the R&D laboratory at GSFC, TAE Plus is also ported and validated with formal acceptance testing on the following UNIX workstations: Apollo, Vaxstation II , Decstation 3100, HP9000, and Macintosh II (A/UX). TAE Plus is also available and validated on the Vaxstation II under VMS and DECWindows™. Other user sites have successfully installed TAE Plus onto the Masscomp, Silicon Graphics Iris and other Unix-based graphic workstations. In January 1991, a beta release of TAE Plus 5.0, which uses the latest version of OSF Motif™ (V1.1), will be available to licensed TAE users on the Sun workstations. In subsequent months, ports to other workstations will become available. There are plans to port TAE Plus to new architectures, including the new IBM 6000 workstation and the 386i class of workstations.

Since the first release of TAE Classic in 1981, we have provided user support through a fully staffed TAE Support Office (TSO). This service has been one of the primary reasons for the success of TAE. Through the TSO, users receive answers to technical questions, report problems, and make suggestions for improvements. In turn, the TSO keeps users up-to-date on new releases, publishes a newsletter, and sponsors user workshops and conferences. This exchange of information enables the Project Office to keep the TAE software and documentation up-to-date and, perhaps most importantly, take advantage of user feedback to help direct future development.

APPLICATIONS USING TAE PLUS

Since 1982 over 750 sites have installed TAE Classic and/or TAE Plus. The applications built or being built with TAE perform a variety of different functions. TAE Classic usage was primarily used for building and managing large scientific data analysis and data base systems (e.g., NASA's Land Analysis System (LAS), Atmospheric and Oceanographic Information Processing System (AOIPS), and JPL's Multimission Image Processing Laboratory (MIPL) system.) Within the NASA community, TAE Plus is also used for scientific analysis applications, but the heaviest concentration of user applications has shifted to support of realtime control and processing applications. This includes supporting satellite data capture and processing, monitor and control of spacecraft and science instruments, prototyping user interface of the Space Station Freedom crew workstations a
rting diagnostic display windows for realtime control systems in ground operations. For thes
applications, TAE Plus is principally used to design and manage the user interface, which is m
f a combination of user entry and data-driven interaction objects. TAE Plus becomes a part of
opment life cycle as projects use TAE Plus to prototype the initial user interface design and have t
ned user interface evolve into the operational UI.

Outside the NASA community, TAE
eing used by an assortment of other government agencies (22%), universities (15%), and private
ies (35%). Within the government sector, users range from the National Center for Atmospheric R
h, National Oceanographic and Atmospheric Adminstration, U.S. Geological and EROS Data Center,
are developing scientific analysis, image mapping and data

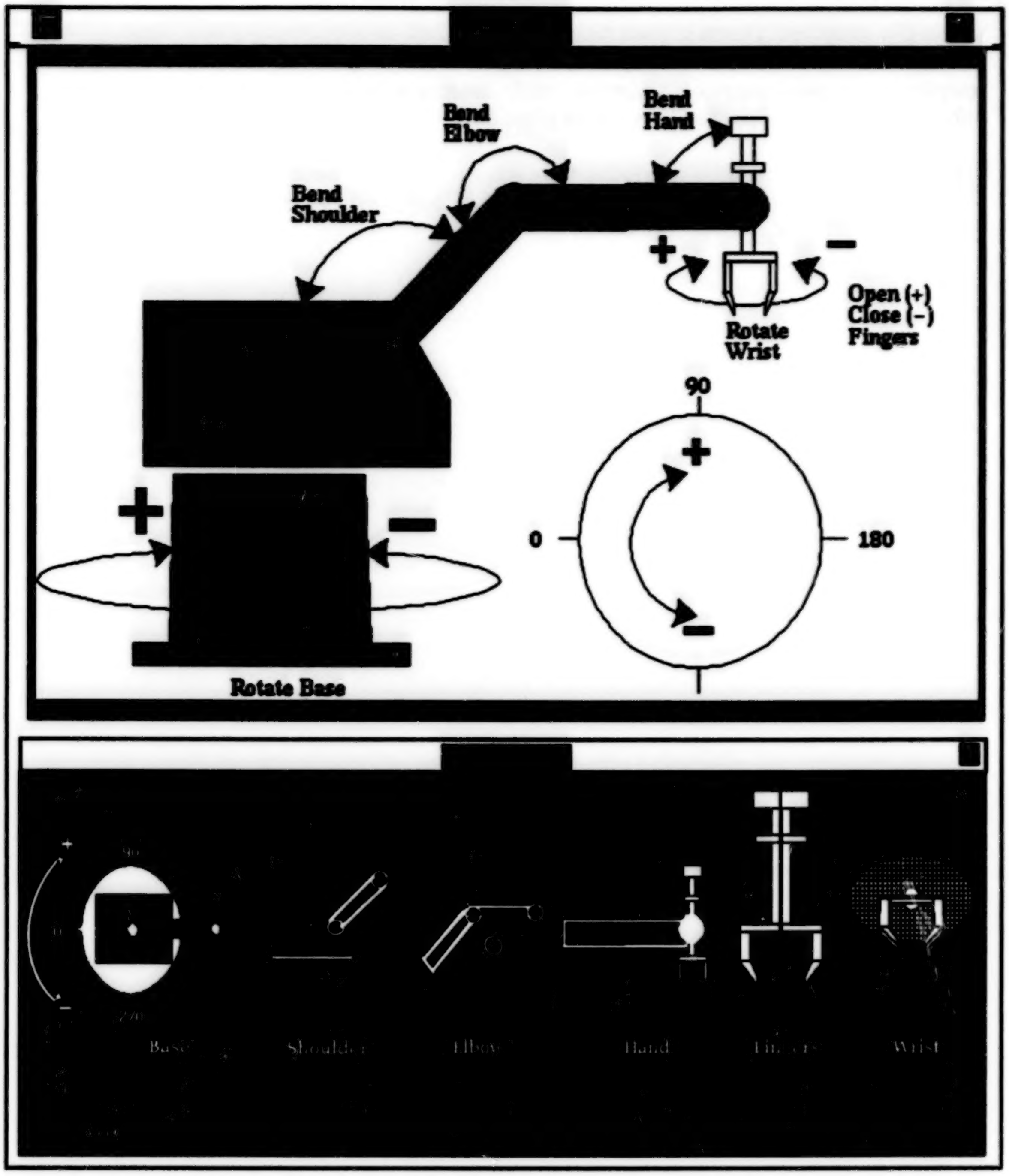

Bend Elbow
Bend Hand
Bend Shoulder
+
−
Open (+) Close (−) Fingers
Rotate Wrist
90
+
0
180
−
+
−
Rotate Base
90
270
Base
Shoulder
Elbow
Hand
Fingers
Wrist

Figure 6.
Robot control panel developed at University of Colorado

distribution systems, to numerous Department of Defense laboratories, who are building command-and-control-related systems. Universities represented among the TAE community include CalTech, Cornell, Georgia Tech, MIT, Stanford, University of Maryland and University of Colorado. Applications being developed by University of Colorado include the Operations and Science Instrument Support System (OASIS), which monitors and controls spacecraft and science instruments and a robotics testbed for research intor the problems of construction and assembly in space. [Ref. 8] Figure 6 shows a view of the robot arm, built with TAE Plus. As the current location of the robot changes, the data-driven objects change respectively. Private industry has been a large consumer of the TAE technology and a sample of the companies that have received TAE Plus V4.1 include Apple Computer Inc., Ford Aerospace, Martin

ORIGINAL PAGE IS
OF POOR QUALITY

Marietta, Computer Sciences Corp., TRW, Lockheed, IBM, Northern Telecom, Mitre Corp., General Dynamics and GTE Government Systems. These companies are using TAE Plus for an assortment of applications, ranging from a front-end for a corporate database to advanced network control center. Northern Telecom, Inc. used TAE Plus to develop a technical assistance service application which enables users to easily access a variety of applications residing on a network of heterogeneous host computers. [Ref. 9] Because of the high cost associated with programming and software-development, more and more software development groups are looking for easy-to-use productivity tools, and TAE Plus is becoming recognized as a viable tool for developing an application's user interface.

NEXT STEPS

The current TAE Plus provides a useful tool within the user interface development environment -- from the initial design phases of a highly interactive prototype to the fully operational application package. However, there are many enhancements and new capabilities that will be added to TAE Plus in future releases.

In the near term, the emphasis will be on enhancement features and upgrades, with the support for the Open Software Foundation's (OSF) Motif™ style and optimizing the TAE Plus software to improve real-time performance being of highest priority. All the requested enhancements are user-driven, based on actual experience using TAE Plus, or requirement-driven based on an application's design. For example, on the enhancements list are extensions to the interaction objects, (e.g., graph data-driven object, form fill-in), support for importing foreign graphics, and refinements in the code generation feature.

Future advancements include expanding the scope of the Transportable Applications Environment (TAE) to include new tools or technologies. For instance, the introduction of hypermedia technology and the integration of expert system technology to aid in making user interface design decisions are targeted for investigation and prototyping.

CONCLUSION

With the emergence of sophisticated graphic workstations and the subsequent demands for highly interactive systems, the user interface becomes more complex and includes multiple window displays, the use of color, graphical objects and icons, and various selection techniques. Prototyping of different user interface designs, thus, becomes an increasingly important method for stabilizing concepts and requirements for an application. At GSFC, the TAE Plus development team had the requirement to provide a tool for prototyping a visual representation of a user interface, as well as to establish an integrated development environment that allows prototyped user interfaces to evolve into operational applications. TAE Plus is fulfilling this role by providing a usable, generalized, portable and maintainable package of development tools.

TAE Plus is an evolving system, and its development will continue to be guided by user-defined requirements. To date, each phase of TAE Plus's evolution has taken into account advances in virtual operating systems, human factors research, command language design, standardization efforts and software portability. With TAE Plus's flexibility and functionality, it can contribute both more advances and more standardization in user interface development system technology.

ACKNOWLEDGEMENTS

TAE Plus is a NASA software product being developed by the NASA/Goddard Space Flight Center with contract support by Century Computing, Inc. The work is sponsored by the NASA Office of Space Operations.

TAE is a registered trademark of National Aeronautics and Space Administration (NASA). It is distributed through NASA's distribution center, COSMIC, (404) 542-3265. For further information, contact COSMIC and/or the TAE Support Office at GSFC, (301) 286-6034.

REFERENCES

1. Perkins, D.C., Howell, D.R., Szczur, M.R., "The Transportable Applications Executive -- an interactive design-to- production development system," *Digital Image Processing In Remote Sensing*, edited by J-P Muller, Taylor & Francis Publishers, London, 1988.

2. Scheifler, Robert W., Gettys, Jim., "The X Window System," MIT Laboratory for Computer Science, Cambridge, MA, October 1986.

3. Open Software Foundation, Inc., *OSF/Motif™ Programmer's Reference Manual*, Revision 1.1, 1990

4. Cox, Brad J., *Object Oriented Programming, An Evolutionary Approach*, Addison-Wesley Publishing Company, Reading, MA, 1986.

5. Stroustrup, Bjarne, *The C++ Programming Language*, Addison-Wesley Publishing Company, Reading, MA, 1987.

6. Szczur, Martha R., Miller, Philip, "Transportable Applications Environment (TAE) Plus: Experiences in 'Object'ively Modernizing a User Interface Environment," Proceedings of the OOPSLA Conference, September 1988.

7. Linton, Mark A., Vlissides, John M., Calder, Paul R., "Composing User Interfaces with Interviews," *IEEE Computer*, February, 1989.

8. Klemp, Marjorie, "TAE Plus in a Command and Control Environment", Proceedings of theTAE Eighth Users' Conference, June, 1990

9. Sharma, Alok, et al., "The TAS Workcenter: An Application Created with TAE", Proceedings of the TAE Eighth Users' Conference, June, 1990

N91

24049

UNCLAS

N91-24049

APPLICATIONS OF FUZZY LOGIC TO CONTROL AND DECISION MAKING

Robert N. Lea
Software Technology Branch / PT4
NASA/Johnson Space Center
Houston, Texas 77058
(713) 483-8085
Nasamail: RLea

Yashvant Jani
LinCom Corporation
Houston, Texas 77058
(713) 488-5700

ABSTRACT

Long-range space missions will require high operational efficiency as well as autonomy to enhance the effectivity of performance. Fuzzy logic technology has been shown to be powerful and robust in interpreting imprecise measurements and generating appropriate control decisions for many space operations. Several applications are underway at the Software Technology Laboratory, Johnson Space Center, investigating the fuzzy logic approach to solving control and decision making problems. Fuzzy logic algorithms for relative motion and attitude control have been developed and demonstrated for proximity operations. Based on this experience, motion control algorithms that include obstacle avoidance were developed for a Mars Rover prototype for maneuvering during the sample collection process. A concept of an intelligent sensor system that can identify objects and track them continuously and learn from its environment is under development to support traffic management and proximity operations around the Space Station Freedom. For safe and reliable operation of Lunar/Mars based crew quarters, high speed controllers with ability to combine imprecise measurements from several sensors will be required. A fuzzy logic approach that utilizes high speed fuzzy hardware chips for such a task is under investigation.

INTRODUCTION

Since its inception by Lotfi Zadeh [1] in the 1960's, fuzzy logic has been applied to many fields [2] including space operations. Applications of fuzzy logic have been developed for the star-tracker data processing system of the Space Shuttle [3], the attitude control [4] and a combined translational and rotational control of a spacecraft [5]. Currently, there is some concentrated effort in the Software Technology Branch of the Information Technology Division at the NASA Johnson Space Center (JSC), directed towards the development of fuzzy logic software capabilities for building expert systems. Particularly, the emphasis has been on developing intelligent control systems for space vehicles and robotics. Also the problem of sensor data monitoring and control of data processing, which includes detection of potential failures in the system and in some cases reconfiguration, has been investigated. Results of the performance tests made on simulated operational scenarios have been very promising. The issues of when, why, and how hardware implementation can be beneficial are also being studied carefully.

There are certain key technology utilization questions to be answered relative to the use of fuzzy logic control over conventional control.

1) Is it possible to create control systems which do not require a high degree of **redesign** when system configurations change or operating environments differ? In other words, can adaptivity be achieved through the use of a fuzzy controller in place of a conventional controller ?

2) Can a fuzzy controller be used as a high level controller to function in cooperation with classical controllers in a way a human would ? Specifically, can it be designed to monitor the system, evaluate its performance, and either suggest or force changes to make the system work properly or at least function more efficiently ?

3) How easy or difficult is it to design and implement a fuzzy rule base that will control a complex system as opposed to developing the classical control system to do the same problem ?

4) Another question of interest particularly to NASA is where can hardware implementations be utilized advantageously and how easy or difficult is it to transfer fuzzy rules to hardware? In many cases hardware will be able to take much of the computational burden off the central computing system.

Fuzzy logic based controllers will be valuable in systems that are highly non-linear and having complex environments that are practically impossible to model. Fuzzy controllers work for linear systems also but probably have less justification in this case, unless the problem is best thought of in a rule-based framework. The Japanese researchers and engineers [6,7,8] have demonstrated the usefulness of fuzzy controllers in the last few years with some impressive applications from a engineering viewpoint, such as the Sendai train controller [9], the air conditioning control system, the camera autofocusing system, the television auto contrast and brightness control, the applications to automobile transmission and braking control, and applications to control of jitter in camera imaging which requires the distinguishing between real motion in the image which is desired and motion of the camera which needs to be filtered out.

The role of fuzzy logic in autonomous space operations is discussed first. A short summary of each of the applications of fuzzy logic so far accomplished in the Software Technology Laboratory (STL) at the JSC is provided next. A summary and references conclude this paper.

ROLE OF FUZZY LOGIC IN AUTONOMOUS SPACE OPERATIONS

The fuzzy logic approach is **simple** to understand and **easy** to implement as a software module. Fuzzy rules provide a framework to implement the **human thinking process** i.e. the rules reflect the human thought process, such as " If the object is Far Left in the FOV then rotate the camera to the left side ". The entire rule base for the controller can be derived as if a human was performing the controlling task. Thus, the knowledge for controlling a process gained through experience can easily be transferred in a software module to achieve the **desired autonomy**.

Fuzzy logic will be useful in proper interpretation of measurements from sensors that are always corrupted by noise and bias. Accuracy of the sensors represent a challenge that is not always surmountable. Fuzzy logic framework can easily handle imprecise measurements, thus helping the integration process. Also sensor systems may perform incorrectly or at least unexpectedly anomalous for a short time. It is necessary to determine this type of behavior and correctly resolve the situation. Processing of uncertain information using **common sense rules** and **natural language** statements is possible in this fuzzy logic framework.

Implementation of fuzzy membership functions, rules and related processing is made easy by tools like the TIL Shell [10] which has a graphics oriented user interface and fuzzy-C compilers [11] that can generate code for the fuzzy chip or C code to integrate with other software modules. There are several commercial products available in the industry that allow easy implementation of knowledge base, rule-base and user interfaces. For autonomous operations, it is easier and useful to implement control decisions through knowledge base and rules so that the **heuristics** and related **experiential knowledge** can be used for a particular situation.

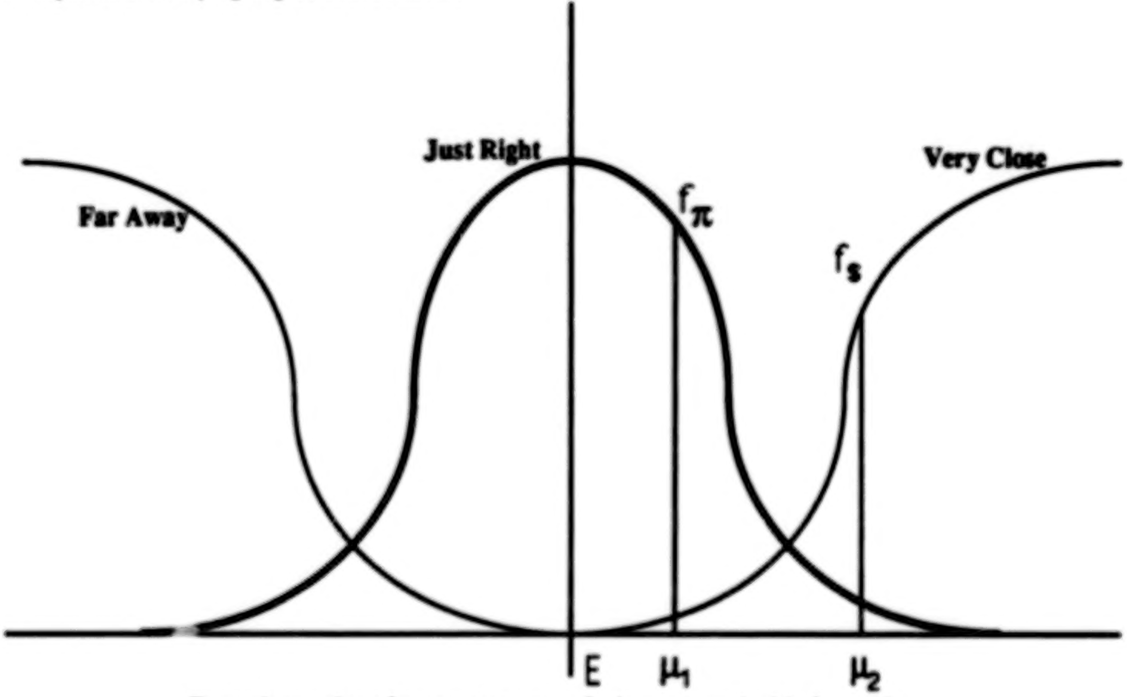
It is also possible to develop and implement a fuzzy controller in the **fuzzy processors**, thus, having a fuzzy hardware controller. There are several commercial fuzzy processors that can process over 30,000 fuzzy rules per second and thus provide high processing power. These fuzzy processors consume low power with a capability to process general purpose instructions and can be mounted in the back plane of a sensor, say, a camera. These processors also provide interfaces to hardware to transfer information and commands to the main Central Processing Unit (CPU). Advanced sensor systems envisioned for space station operations will have such processors embedded as an integral part of the system. Thus, a **distributed processing** function onboard the spacecraft is possible via fuzzy chips.

A camera tracking system described later can be a dedicated sensor with built-in intelligence and speed to perform functions which are normally performed in the main CPU onboard the Space Station Freedom (SSF). With a dedicated fuzzy chip and its processing power, there will be virtually no computational load on the SSF main computers. As a result, the main CPU will be available for other computing requirements such as complex guidance and navigation schemes. Furthermore, the interfaces between the fuzzy chip and the main CPU will be at a command level requiring reasonably low speed data transfer. There will be no need for a high rate data transfer which can increase the cost and decrease the reliability.

A significant application of fuzzy logic is in an advisory role in health monitoring and internal reconfiguration of spacecraft subsystems. These processes require a capability to handle uncertain measurements, estimate possibilities of failures and quickly rearrange flow so that the autonomous operations are not stopped. Techniques have been developed to update the rule base using reinforcement learning [12] in a given environment and adjust the response or behavior of a controller. These are very important for achieving **operational efficiency** in space operations.

RELATIVE MOTION AND ATTITUDE CONTROLLER

Fuzzy sets have been used in developing a trajectory controller for a spacecraft to maintain proximity operations profiles [13,14]. An automated vehicle controller that interprets the sensor measurements in a manner similar to a human expert has been modeled using fuzzy sets. The control rules were derived from the thinking process used by pilots and were implemented using typical pi- and s-functions (fig. 1) that can be adjusted for varying degree of **fuzziness**.



For a given value of parameter, m, f_π is compared with f_s, and the larger value dictates whether action is taken.

Fig. 1 Control Strategy for Range Parameter

The controller has been implemented into a multi-vehicle dynamical simulator known as the Orbital Operations Simulator (OOS) [15], complete with all environment and sensor models. A small part of this control simulation was demonstrated via tele-video links [16], to the International Fuzzy Systems Association (IFSA) Workshop that was held in Iizuka, Fukuoka, Japan in August 88. In this simulation, the automated fuzzy controller was used to control the closing rates and relative positions of the shuttle with respect to the solar max satellite. The scenario required it to perform operations including approach to target, fly around and stationkeeping.

Many different scenarios have been run with this automated fuzzy controller to evaluate the performance with respect to flight profiles and delta-v requirements. Comparisons of delta-v requirements for a man-in-the-loop versus the automated controller have shown that the automated controller always uses less delta-v.

For a test case involving stationkeeping at 150 feet for 30 minutes, the automated controller required 0.1 ft/sec delta-v whereas 0.54 ft/sec was used in the man-in-the-loop simulation. For v-bar approach from 500 feet to 40 feet within a 25 minute time interval, the automated controller used 2.12 ft/sec vs. 2.99 ft/sec for the man-in-the-loop simulation.

To complement this translational control, it was decided to implement the rotational control via fuzzy membership functions and the rules based on the conventional phase plane. It was obvious that such an implementation would provide a direct performance comparison with the conventional control system, thus leading to further insight into understanding the relative merits of fuzzy control systems. Furthermore, an integrated six Degree Of Freedom (DOF) controller can be developed by combining these two control systems.

The rotational control system is based on the phase plane construct used in the attitude control system. The angle and rate errors, PHI and PHI_DOT, are input and torque is the output for this rotational controller. The input variables have seven membership functions defined over the universe of discourse while the output variable torque has five membership functions (fig. 2). The membership functions are piece-wise linearly defined and have graphs that form mostly triangular or trapezoidal shapes. There are 25 rules defined for reducing the PHI and PHI_DOT errors to within their zero (ZO) range.

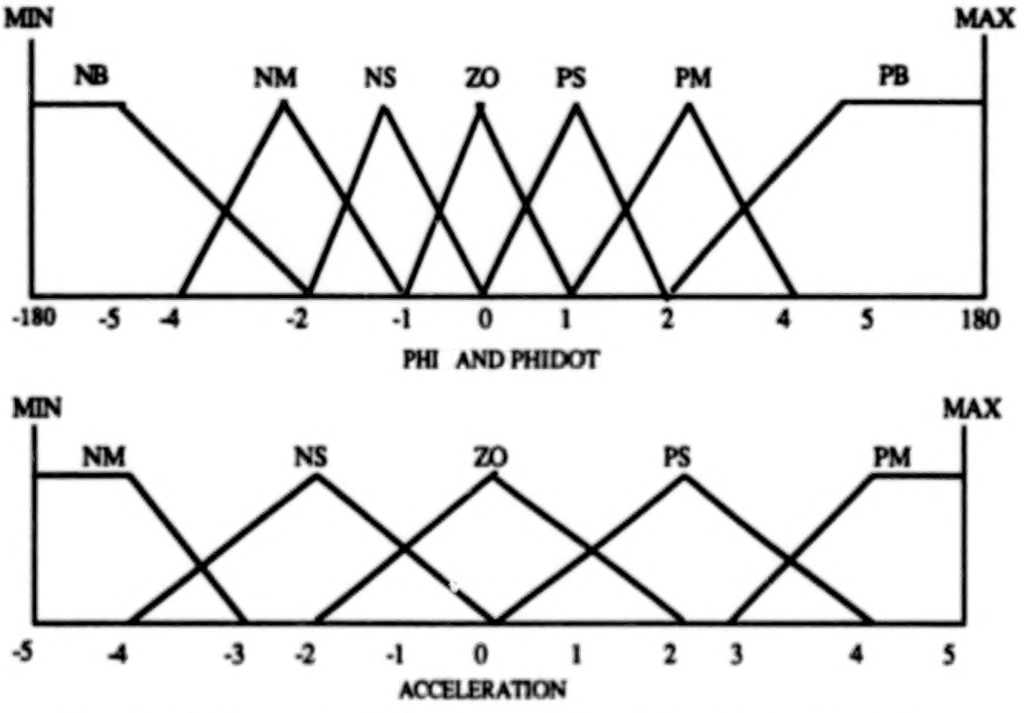


Fig. 2 Membership functions for input and output variables for attitude controller

Single axis rotational equations were implemented for the pitch axis of the shuttle. The pitch moment of inertia and the positive and negative pitch torques provided by jets were used in this simple simulation to test the fuzzy controller rules. The shuttle jets provide a larger acceleration for positive pitch as compared to the negative pitch. The simulation was set up to provide a constant torque during a cycle time of 80 milliseconds. The pitch attitude and the rate are propagated at this cycle time. When the fuzzy controller asks for a torque greater than 0.5, the constant torque is provided in that direction, otherwise no torque is provided. This simulates the jet on and off activity at the appropriate time. Testing for the pitch axis has shown very satisfactory results. With several starting states, meaning initial angle and rate, the system has converged on the commanded value, and manifested relatively smooth limit cycles around the deadband. The control system response in all cases has been as expected, including overshoot behavior in cases where initial rate error is very large. Tests were performed with some rules turned off or deactivated to observe the performance with a limited ruleset. The objective was to reduce the number of rules to a minimum.

Performance of the fuzzy controller with 25 rules (see Table 1.) was more than adequate for a single axis, and gave us confidence to expand it to three axes case. There is an automatic attitude control system called the Digital Auto Pilot (DAP) for the space shuttle. This onorbit DAP is implemented in OOS for shuttle on-orbit operations. There is a module called Phase Plane which is replaced by this rotational fuzzy controller with all other interfaces maintained intact. The integration process was completed with only minor modifications to the interfaces. The simulation testing included three axes attitude hold and single axis maneuvers. In a three axes attitude hold case, the fuzzy logic based controller used only 30 % of the fuel used by the DAP, while for an attitude maneuvers case, the fuzzy controller used around 60 % of the fuel used by the DAP. In both cases, the fuzzy controller has shown comparable performance for maintaining attitude and body rates. Further testing and analysis is planned to include other maneuver modes and different parameters sets.

Table 1. Rule base for attitude controller

	phi						
phi_dot	NB	NM	NS	ZO	PS	PM	PB
NB	PM	PM	PS				
NM	PM	PM	PS				
NS	PS	PS	PS				
ZO	PS	PS	ZO	ZO	ZO	NS	NS
PS					NS	NS	NS
PM					NS	NM	NM
PB					NS	NM	NM

The integration approach adopted for combining translational and rotational control systems is simple, straight forward and involves extensive testing. The translational fuzzy control system will be used by the autosequencer to generate proper hand controller commands so that the desired range and range rate are maintained during proximity operations. Typically, a shuttle pilot provides these inputs and controls the relative trajectory. Thus the autosequencer will simulate the crew input via the translational fuzzy control system. The rotational fuzzy control system as described earlier will generate commands for jet-select to fire jets for attitude control. Existing interfaces with the Phase Plane module will be maintained intact for the overall integrity of the system. When both fuzzy control systems are used together, it will provide a total 6 DOF controller for proximity operations.

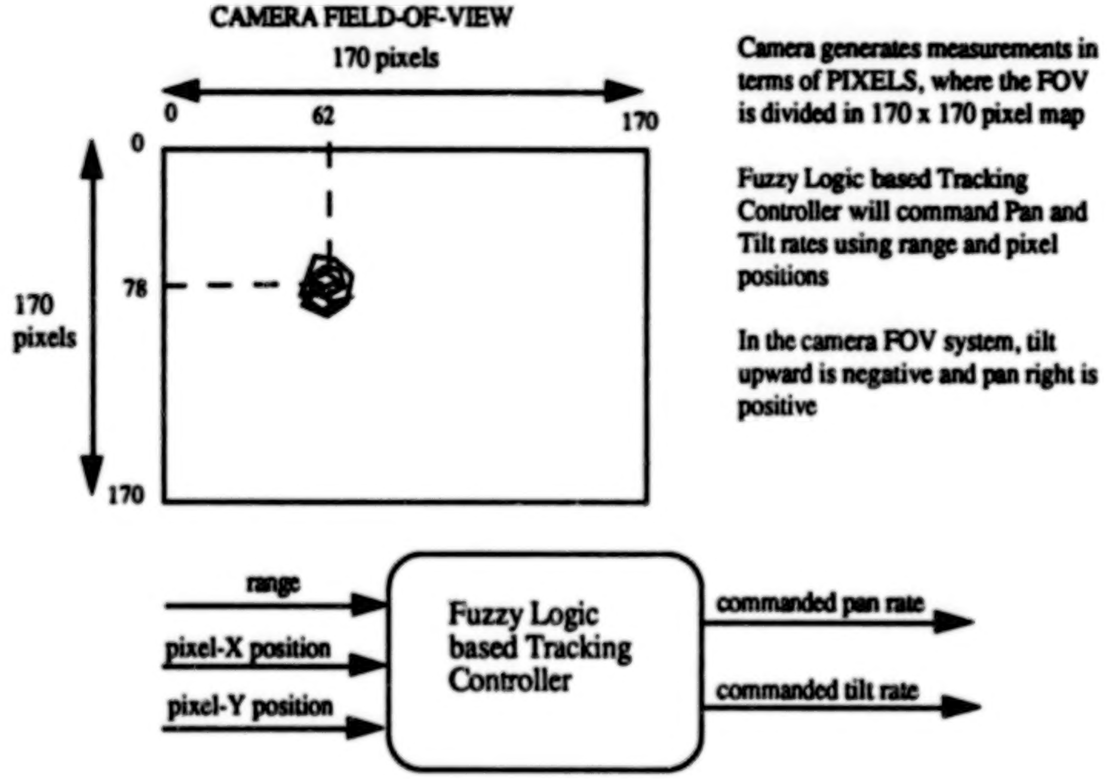
A preliminary test plan has been put together to test the 6 DOF controller. It includes test cases for stationkeeping with a fixed attitude, stationkeeping with attitude changes, line of sight approach on the V-bar, line of sight approach on the R-bar, fly around at a constant distance with constant relative attitude, and final approach for docking. Details of these test cases such as initial conditions, commanded attitude maneuvers, etc. are being defined to finalize the test plan.

INTELLIGENT SENSOR SYSTEM

Advanced sensor systems with intelligence and a distributed nature will be required for activities like proximity operations and traffic control around the SSF. There will be several sensors of different types providing various measurements simultaneously as inputs for processing to such a system. The conceptual development of such a system where several cameras, laser range finders and radar can be used as independent components is in progress within the Software Technology Laboratory of the Information Technology Division at the Johnson Space Center. The first phase of this development is the camera tracking system based on the fuzzy logic approach that utilizes the object's pixel position as inputs and controls the gimble drives to keep this object in the Field Of View (FOV) of the camera.

Tracking of an object means aligning the pointing axis of a camera along the object's line of sight. The monitoring camera is typically mounted on the pan and tilt gimble drives which are capable of rotating the pointing axis within a certain range. The task of the tracking controller is to command these gimble drives so that the pointing axis of the camera is along the line of sight vector which is estimated from the measurements.

For the fuzzy logic based tracking controller, the inputs are range and line of sight vector, and the outputs are the commanded pan and tilt rates (fig. 3). The line of sight vector is input in terms of pixel position in the camera FOV. When an image is received, it is processed to determine the location of the object in the camera frame which has the vertical, horizontal and pointing vectors as three axes. Usually an image, particularly for complex objects, spans over many pixels. Using a suitable technique, the centroid of the image is computed as the current location in the viewing plane which is like a Cartesian coordinate plane having vertical and horizontal axes. The size of the viewing plane is 170 x 170 pixels, and the origin is at the upper left corner as shown in fig. 3. The range of the object is received from the laser range finder as a measurement. These three parameter values are input to the controller. There are five membership functions for horizontal and vertical positions as well as range input [17].


CAMERA FIELD-OF-VIEW

Camera generates measurements in terms of PIXELS, where the FOV is divided in 170 x 170 pixel map

Fuzzy Logic based Tracking Controller will command Pan and Tilt rates using range and pixel positions

In the camera FOV system, tilt upward is negative and pan right is positive

Fig. 3 CONCEPT OF A CAMERA TRACKING SYSTEM

The desired image location is the center of the viewing plane, which is at (85,85). If the current location is close to the center, then rotation of the pointing axis is not required. If the location is to the left of center then a left rotation is necessary. Similarly, if the image is down from the horizontal line then a downward rotation is required. These rotations are determined using the position and range measurements and the rule base shown in Table 2. First the range measurement is fuzzified and the value of the scale factor is determined based on the scale_factor rules. Necessary defuzzification processing is performed to compute the crisp value of the scale factor. Then, the scale factor and the position measurements are provided to the next set of rules to determine the rate at which the gimble drives should be rotated. There are 30 rules that determine both pan and tilt rates. Again, the necessary defuzzification processing is performed to compute the crisp values of the pan and tilt rates which can be sent to the gimble drives as command values.

Table 2. Rule base for the tracking task

	Distance Membership Functions				
	VFAR	FAR	NEAR	VNEAR	PROX
Scale_Factor	LOW	LOW	MED	HIGH	HIGH

		Horizontal Position Membership Functions				
		FL	LL	CENTER	LR	FR
Scale_Factor	LOW	FN	SN	ZR	SP	FP
	MED	SN	SN	ZR	SP	SP
	HIGH	SN	ZR	ZR	ZR	SP
		Pan_Rate Membership Functions				

		Vertical Position Membership Functions				
		FD	LD	CENTER	LU	FU
Scale_Factor	LOW	FP	SP	ZR	SN	FN
	MED	SP	SP	ZR	SN	SN
	HIGH	SP	ZR	ZR	ZR	SN
		Tilt_Rate Membership Functions				

Note - Negative Tilt_rate means the pointing axis going upward in FOV

The camera is moved based on these commands within the limits of its gimble rates and angles. New measurements in the camera FOV are obtained for the next cycle and the processing is repeated. The cycle time is based on the processing time required for the following functions : 1) determining pixel positions, 2) obtaining a range measurement, 3) rotating the gimble drives at a desired rate, and 4) the requirements to track the object within a certain accuracy. Typical cycle time ranges between 0.1 to 1.0 second.

There are several advantages of our approach that utilizes fuzzy logic in a camera tracking system. This system will be a low power sensor as compared to an active sensor e.g. Radar in the Ku band range, or LADAR using laser frequency. Typically, the active sensor radiates a power pulse towards a target and receives back a reflected pulse. Based on the power transmitted, power received and time between these pulses, parameters like range and range rates are calculated. Since the camera tracking system will not be radiating power, it will be a low power sensor in comparison with an active sensor. Since there is already a shortage of power, an important consumable, onboard the SSF, availability of low power sensors is very important for continuous operations. The SSF can afford to keep this type of a sensor working around the clock without having much impact on the power management or other computational load on the main computers.

Capabilities of the tracking controller can be expanded to perform other functions such as approach toward the object, grapple, object identification, traffic management, and caution and warning to the crew. Fast moving objects can be identified easily via prediction of position and thus collision avoidance can also be achieved. Since the system can work as a stand-alone system at the command level and will interrupt the operations flow only if necessary, it can become a node in the distributed sensor system.

Current plans include testing of the concept in software simulations in the STL and in the hardware laboratories in Engineering Directorate at JSC. The software testing will refine the Rule Base and the Membership Functions, while the hardware testing will identify all interface problems, real-time performance evaluation, and the controller behavior in light of actual measurements which will be noisy and imprecise. Both type of testing is required in order to make the system operational and useful. Development of pattern recognition and object identification algorithms is underway in the STL [18].

MARS ROVER TRAJECTORY CONTROL AND PLANNING

While collecting soil samples and surveying the Mars surface, the Mars rover will be moving from one point to another among obstacles which cannot be identified prior to the mission. In order to complete the collection task, the rover must interpret imprecise sensor measurements of obstacle size and distance to determine which obstacles present a hazard and must be avoided and plan a trajectory to avoid these unforeseen obstacles. In addition, worst case round trip communications time between Earth and Mars will require 20 minutes. Earth-based tele-robotic control of the Mars rover will be extremely difficult and time consuming and could seriously endanger the success of the mission. Fuzzy trajectory planning and control provides robust real-time control capable of adapting the trajectory profile to avoid unforeseen hazards. The fuzzy logic approach eliminates communications travel time, allows the rover to avoid obstacles which may be unavoidable due to tele-robotic reaction time, and provides adaptable control which will extend the rover performance envelope.

A fuzzy logic approach to trajectory control has been developed which allows the rover to avoid these hazards during the sample collection process. The fuzzy trajectory controller receives the goal or target point from the planner and uses X and Y position errors as well as Orientation (Yaw) error in the control system frame and commands the rover in terms of steering angle and velocity. The fuzzy rule-base containing 112 rules for the controller has been designed to drive the rover towards the X-axis of control error frame. As the rover approaches this axis, the rover is commanded to the correct orientation error and then slowly drive towards the target point.

The X and Y position error variables were modeled as a shouldered [11] membership set of 5 piece-wise linear functions with a universe of discourse ranging from -100 to 100 meters. The orientation or yaw error variable was modeled as an unshouldered membership set of 7 functions with a universe of discourse ranging from -180 to 180 degrees. The steering variable was modeled as an unshouldered membership set of 5 functions with a universe of discourse ranging from -30 to 30 degrees. Finally, the velocity variable was modeled as an unshouldered membership set of 7 functions with a universe of discourse ranging from -5 to 5 meters/second.

A fuzzy trajectory controller for a Mars rover has been tested on several cases. Preliminary results have shown that the trajectory controller can reach the target position and attitude within 0.0005 meters on the x-error axis, 0.25 meters on the y-error axis, and 0.45 degrees yaw error. It is believed that these accuracies can be reduced by altering the membership function sets for the inputs and outputs. Further testing will facilitate the tailoring of the membership functions to the fuzzy rule set. The fuzzy approach provides a control system which can be easily modified and tested.

CONCEPT FOR LUNAR/MARS CREW QUARTERS CONTROL

Continuous monitoring of the Environment and Life Support System (ELSS) for Lunar/Mars crew quarters will be required for two reasons; 1) the safety of the crew, and 2) an efficient usage and management of available resources. The system dynamics model (typically known as the 'plant' in conventional control theory) that represents the behavior of the system becomes increasingly complex and non-linear as the volume of the crew quarters increases significantly. Multiple sensor measurements distributed over the entire volume are required to derive accurate state information for the system so that a nominal operational state can be maintained. In such a case, applying conventional control theory will be very difficult, if not impossible. A concept of a fuzzy logic based monitoring and diagnosis technique is under development to combine several measurements from different types of sensors and maintain a desired state of this non-linear system. The concept can easily be expanded to detect potential component failures and generate immediate advisory messages for corrective actions. Suitability of currently available fuzzy hardware for real-time monitoring and diagnosis is also being investigated.

SUMMARY

Applications of fuzzy logic in autonomous orbital operations are described in this paper with past accomplishments at JSC. Current ongoing as well as future activities planned are also described. The main objective of all these activities is to **increase autonomy** in orbital operations and thus achieve a higher level of **operational efficiency** desired for future space operations. The approach is to develop modular control that can be upscaled for greater autonomy in an integrated environment. The initial step is to develop a software controller and then to integrate it with hardware at the appropriate level. As the activities progress, detail testing is performed to check out implementation and integration of components. Our preliminary results promise a very successful utilization of fuzzy logic in autonomous orbital operations.

REFERENCES

1. Zadeh, L. : "Fuzzy Sets", Information and Control, vol. 8, pp. 338-353, 1965.
2. Klir G. J. ; and Folger T. A. : Fuzzy sets, Uncertainty, and Information, Prentice Hall, New Jersey, 1988.
3. Lea, R. N. ; and Giarratano, J. : An Expert System Program Using Fuzzy Logic For Shuttle Rendezvous Sensor Control, proceedings of ROBEXS'86, pp. 327-329, 1986.
4. Lea, R. N. ; and Jani, Y. : Spacecraft Attitude Control System Based on Fuzzy Logic Principles, proceedings of ROBEXS'89, 1989.
5. Lea, R. N. ; Togai, M. ; Teichrow, J. ; and Jani, Y. : Fuzzy Logic Approach to Combined Translational and Rotational Control of a Spacecraft in Proximity of the Space Station, Proceedings of the IFSA'89, pp. 23-29 1989.
6. Rogers, M. ; and Hoshiai, Y. : The Future Looks 'Fuzzy', NEWSWEEK, May 28, 1990.
7. Johnson, R. C. : Clear Leader Emerges : Japan at fuzzy fore, EETimes, Sept. 11, 1989.
8. Armstrong, L. ; and Gross, N. : Why 'Fuzzy Logic' beats black-or-white thinking, Science & Technology section, BUSINESS WEEK, May 21, 1990.
9. Yasunobu, S. ; and Miyamoto, S. : "Automatic Train Operation System by Predictive Control", Industrial Applications of Fuzzy Control, Sugeno, M. (Ed.), 1-18, North-Holland: Amsterdam, 1985.
10. Perkins, C. ; Teichrow, J. ; and Horstkotte, E. : Fuzzy-C development system : A complete overview, Togai InfraLogic Inc., SOAR-89 conference held at Johnson Space Center, Houston, July 25-27, 1989.
11. Teichrow, J. ; and Horstkotte, E. : Fuzzy-C compiler User's manual, v2.0b, Togai InfraLogic Inc., Irvine, California, April 1989.
12. Lee, C. C. ; and Berenji, H. R. : An Intelligent Controller Based On Approximate Reasoning And Reinforcement Learning, Proc. of IEEE Int. Symposium on Intelligent Control, Albeny, NY 1989.
13. Lea, R. N. : Automated Space Vehicle Control for Rendezvous Proximity Operations, Telematics and Informatics, vol. 5, no. 3, pp 179-185, 1988.
14. Lea, R. N. : Applications of fuzzy sets to Rule-based Expert System Development, Telematics and Informatics, vol. 6, nos. 3/4, pp 403-406, 1989.
15. Edwards, H. C. ; and Bailey, R. : The Orbital Operations Simulator User's Guide, LinCom corporation, ref. LM85-1001-01, June 87.
16. Video Conference Demonstration from Johnson Space Center, International Workshop On Fuzzy Systems Applications (IFSA-88), Iizuka, Fukuoka, Japan, August 20-24 1988.
17. Lea, R. N., Giarratano, J., Fritz, R. H., and Jani, Y. K. : Fuzzy Logic Control for Camera Tracking System, Proceedings of the 8th International Congress of Cybernetics and Systems, New York, June 1990.
18. Pal, S. K. : 'Fuzziness, Image Information and Scene Analysis' in An Introduction to Fuzzy Logic Applications in Intelligent Systems edited by R. R. Yager & L. A. Zadeh, Kluwer Academic Publishers (to appear).

N91

24050

UNCLAS

N91-24050

GENETIC ALGORITHMS

Lui Wang
Software Technology Branch
Lyndon B. Johnson Space Center
National Aeronautics and Space Administration
Houston, TX 77058
wang@mpad.span.nasa.gov

Steven E. Bayer
Engineering Technology Group
The MITRE Corporation
1120 NASA Road One, Suite 600
Houston, TX 77058
bayer@mpad.span.nasa.gov

ABSTRACT

Genetic algorithms are highly parallel, mathematical, adaptive search procedures (i.e., problem-solving methods) based loosely on the processes of natural genetics and Darwinian survival of the fittest. This paper introduces basic genetic algorithm concepts, discusses genetic algorithm applications, and presents results from a project to develop a software tool that will enable the widespread use of genetic algorithm technology.

INTRODUCTION

Background

Genetic algorithms (GAs) were pioneered by John Holland in his research on adaptation in natural and artificial systems (1). This research outlined a logical theory of adaptive systems. In essence, biological adaptive systems strive to optimize single individuals or entire species for specific environments to increase the chance of survival. Holland simulated the methods used when biological systems adapt to their environment in computer software models–the genetic algorithms–to solve optimization and machine learning problems. The following paragraphs briefly discuss two types of adaptation strategies which are observed in many biological systems and inspired the basic framework of genetic algorithms.

Adaptation. One form of adaptation pertains to the way an individual changes within its environment to promote survival. Examples include the development of antibodies specific to certain diseases, or the enlargement of muscles needed for daily activities. The way we learn, and the neural changes that accompany learning, is another example of how an individual adapts within its environment. The effects of this form of adaptation are not imprinted on the genome (the genetic makeup of a species); that is, they are not passed on from generation to generation. On the other hand, *individual* adaptation does promote the survival of the individual within an environment–*survival of the fittest*–and enhances that individual's net reproductive advantage through a *natural selection* where *fitter* members of a population are more likely to reproduce.

All species have used adaptive search for millions of years, through an evolutionary search process, to improve the way a species lives and survives within its environment. Therefore, adaptation also refers to evolution and modification of an entire species to fit its environment. This is the process of making a species environmentally *fit*. An appropriate example can be seen in the way many plant species have evolved their flower to resemble a female bee or wasp that attracts the male counterpart and promotes pollination. This evolutionary or *species* adaptation is imprinted on the genome and is passed on to subsequent generations.

Thus natural, biological systems continuously use adaptive search to improve genomes–that is, to improve the species–and to promote the survival of *fitter* individuals and genomes through natural selection.

Genetic Algorithms. Genetic algorithms are highly parallel, mathematical, adaptive search procedures (i.e., problem-solving methods) based loosely on the processes of natural genetics and Darwinian survival of the fittest. These algorithms apply genetically-inspired operators to populations of potential solutions in an iterative fashion, creating new populations while searching for an optimal (or near-optimal) solution to the problem at hand. Population is a key word here: the fact that many points in the space are searched in parallel sets genetic algorithms apart from other search operators. Another important characteristic of genetic algorithms is the fact that they are very effective when searching (e.g., optimizing) function spaces that are not smooth or, continuous–functions which are very difficult (or impossible) to search using calculus based methods. Genetic algorithms are also blind: that is, they know nothing of the problem being solved other than payoff or penalty (i.e., objective function) information.

The basic iterative model of the genetic algorithms is shown in figure 1. A new population is created from an existing population by means of *evaluation*, *selection*, and *reproduction*. This process repeats itself until the population converges on an optimal solution or some other stopping condition is reached.

The initial population consists a set of individuals (i.e., potential solutions) generated randomly or heuristically. In the classical genetic algorithm, each member is represented by a fixed-length binary string of bits (a *chromosome*) that encodes parameters of the problem. This encoded string can be decoded to give the integer values for these parameters.

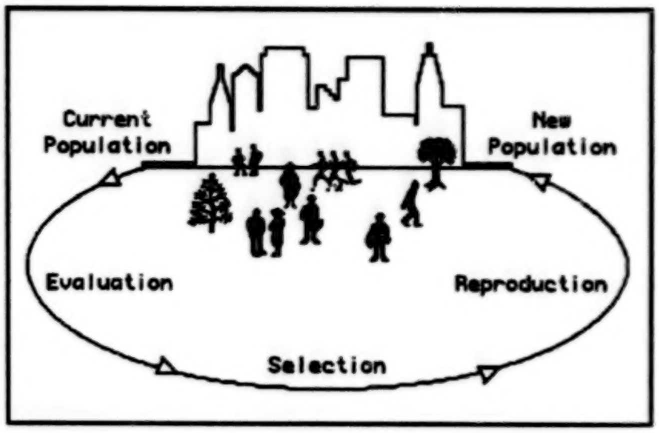


Figure 1. The Iterative Genetic Algorithm Model.

Once the initial population has been created, the evaluation phase begins. The genetic algorithms require that members of the population can be differentiated according to *goodness* or *fitness*. The members that are more fit are given a higher probability of participating during the selection and reproduction phases. Fitness is measured by decoding a chromosome and using the decoded parameters as input to the objective function. The value returned by the objective function (or some transformation of it) is used as the fitness value.

During the selection phase, the population members are given a target sampling rate which is based on fitness and determines how many times a member will mate during this generation–that is, how many offspring from this individual will be created in the next population. The target sampling rate (usually not a whole number) must be transformed into an integer number of matings for each individual. There are many ways of determining the target sampling rate and the actual number of matings. Suffice it to say that individuals that are more fit are given a reproductive advantage over less fit members.

During the reproduction phase, two members of the mating pool (i.e., members of the population with non-zero mating counts) are chosen from random and genetic operators are applied to their genetic material to produce two new members for the next population. This process is repeated until the next population is filled. The recombination phase usually involves two operators: crossover and mutation. A simple crossover operation is illustrated in figure 2. During crossover, the two parents exchange substring information (genetic material) at a random position in the chromosomes to produce two new strings.

Crossover occurs according to a crossover probability, usually between 0.5 and 1.0. The crossover operation searches for better *building blocks* within the genetic material which combine to create optimal or near-optimal problem parameters and, therefore, problem solutions, when the string is decoded.

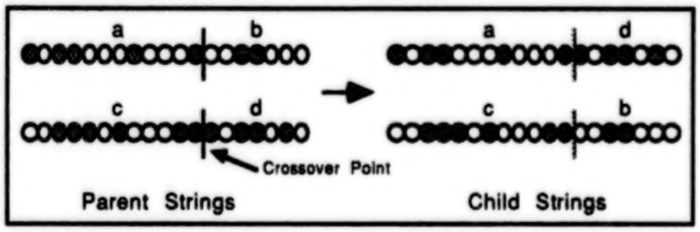


Figure 2. The Crossover Operation.

The mutation operation, shown in figure 3, is a secondary genetic algorithm. It is used to maintain deversity in the population–that is, to keep the population from prematurely converging on one solution–and to create genetic material that may not be present in the current population. The mechanics of the mutation operation are simple: for each position in a string created during crossover, change the value at that position according to a mutation probability. The mutation probability is usually very low–less than 0.05.

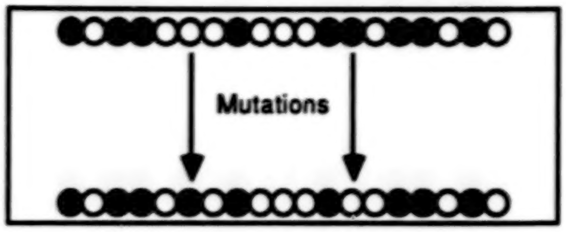


Figure 3. The Mutation Operation.

Genetic Algorithm Applications

Since genetic algorithms provide a set of efficient, domain-independent search methods, they have been used for a wide range of applications. Table 1 lists some of the GA applications ranging from science and engineering to business and social sciences applications. The following sections briefly describe several of these applications.

Engineering. Goldberg (3) applied genetic algorithms and classifier systems to optimization and machine learning problems in natural gas pipeline control. He focussed on a 10-compressor, 10-pipe, steady-state, serial pipeline problem. The object was to minimize the power consumed subject to maximum and minimum pressure and pressure ratio constraints. Goldberg and Samtani (4) used a simple genetic algorithm to optimize a 10-member plane truss. The objective was to minimize the weight of the structure subject to maximum and minimum stress constraints on each member. In both cases, optimal or near-optimal results were obtained.

Medical Image Processing. Fitzpatrick, et al. (5), used genetic algorithms to perform image registration for an arterial examination system known as digital subtraction angiography. Using this system, a physician examines arteries using two x-rays: one taken of the artery unaltered and one taken after injection of a dye into the artery. The two x-ray images are subtracted one pixel at a time; the result is a picture of the interior wall of the artery. Movement of the artery between the time each x-ray is taken results in a distorted image. Fitzpatrick, et al., used genetic algorithms to find a set of equations that transform or *register* the two images.

Robot Path Planning. A Mobile Transporter system is being designed for on-orbit use with Space Station Freedom which will be capable of traversing the station's truss structure. The Mobile Transporter's function is to facilitate space station maintenance tasks and transportation of material around the station. The Software Technology Branch has investigated the use of genetic algorithms for Mobile Transporter path planning (6). The objectives of these activities are to produce an optimum trajectory for the Mobile Transporter that avoids collisions with objects attached to the truss and to minimize the length of the path between the Mobile Transporter and the target position.

Machine Learning. Genetic algorithms have been used in an area of machine learning called *classifier systems*. Classifier systems learn if-then production rules that guide the performance of a production system. Holland has used classifier systems in studies of economic models, specifically mathematical stock market models. The genetic algorithm creates new rules for trading and selling stocks.

TABLE 1. Genetic Algorithm Applications in Search and Optimization [taken from (2)]

Year	Investigators	Description
	BIOLOGY	
1967	Rosenberg	Simulation of the evolution of single-celled organism populations
1970	Weinberg	Outline of cell population simulation including metalevel GA
1984	Perry	Investigation of niche theory and speciation with GAs
1986	Grosso	Simulation of diploid GA with explicit subpopulations and migration
1987	Sannier and Goodman	GA adapts structures responding to spatial and temporal food availability

TABLE 1. (Continued.)

Year	Investigators	Description
	COMPUTER SCIENCE	
1967	Bagley	Parameter search in hexapawn-like game evaluation function via GA
1979	Raghavan and Birchard	GA-based clustering algorithm
1982	Gerardy	Probabilities automation identification attempt via GA
1984	Gordon	Adaptive document description using GA
1985	Rendell	GA search for game evaluation function
1987	Raghavan and Agarwal	Adaptive document clustering using GAs
	ENGINEERING & OPERATIONS RESEARCH	
1981	Goldberg	Mass-spring-dashpot system identification with simple GA
1982	Etter, Hicks, and Cho	Recursive adaptive filter design using a simple GA
1983	Goldberg	Steady-state and transient optimization of gas pipeline using GA
1985	Davis	Bin-packing and graph-coloring problems via GA
1985	Davis	Outline of job shop scheduling procedure via GA
1985	Davis and Smith	VLSI circuit layout via GA
1985	Fourman	VLSI layout compaction with GA
1985	Goldberg and Kuo	On-off, steady-state optimization of oil pump-pipeline system with GA
1986	Glover	Keyboard configuration design using a GA
1986	Goldberg and Samtani	Structural optimization (plane truss) via
1986	Goldberg and Smith	Blind knapsack problem with simple GA
1986	Minga	Aircraft landing strut weight optimization with GA
1987	Davis and Coombs	Communications network link size optimization using GA
1987	Davis and Rittter	Classroom scheduling via simulated annealing with metalevel GA
	IMAGE PROCESSING & PATTERN RECOGNITION	
1970	Cavicchio	Selection of detectors for binary pattern recognition
1984	Fitzpatrick, et al.	Image registration via GA to minimize image differences
1985	Englander	Selection of detectors for known image classification
1985	Gillies	Search for image feature detectors via GA
1987	Stadnyk	Explicit pattern class recognition using partial matching
	PHYSICAL SCIENCES	
1985	Shaefer	Nonlinear equation solving with GA for fitting potential surfaces
	SOCIAL SCIENCES	
1979	Reynolds	GA-like adaptation in model of prehistoric hunter-gatherer behavior
1981	Smith and De Jong	Calibration of population migration model using GA search
1985	Axelrod	Simulation of the evolution of behavioral norms with GA
1985	Axelrod	Iterated prisoners dilemma problem solution using GA

THE SPLICER PROJECT

This section introduces the Splicer Project. It presents background material and discusses the objectives of the project, the approach taken, results to date, current status and possibilities for future work.

Background

The Splicer Project is a project within the Software Technology Branch at NASA's Johnson Space Center. The purpose of the project is to develop a tool that will enable the widespread use of genetic algorithm technology.

The charter of the Software Technology Branch is to develop and/or acquire generic software tools for emerging technologies. Genetic algorithms are just one of the many technologies being investigated within the Software Technology Branch: other areas and tools are expert systems (CLIPS), neural networks (NETS), fuzzy logic, scheduling (COMPASS), software reuse, and intelligent computer-aided training.

The MITRE Corporation supports the Software Technology Branch on multiple projects and is responsible for evaluating the viability and robustness of genetic algorithms and for supporting the Software Technology Branch with respect to the development and acquisition of software tools related to this technology.

Objectives

The Software Technology Branch is interested in applying genetic algorithms within various domains: e.g., robot path planning, job shop scheduling. The original goal of the Splicer Project was to create a flexible, *generic* tool. As such, the tool would:

- Implement the basic genetic algorithms defined in the literature
- Define the interfaces for and allow users to develop interchangeable fitness modules
- Provide a graphic, event-driven user interface

Subsequent goals include the following:

- Distribution of the tool in the public domain
- Support for multiple computing platforms
- Extension of the tool for additional genetic algorithm functionality
- Use of the tool for genetic algorithm research
- Augmentation of the tool and special user interfaces for specific application domains

Approach

Design. The design chosen for the Splicer tool is shown in figure 4. This design consists of four components: a genetic algorithm kernel and three types of interchangeable libraries or modules: representation libraries, fitness modules, and user interface libraries.

A *genetic algorithm kernel* was developed that is independent of representation (i.e., problem encoding), fitness function, or user interface type. The GA kernel comprises all functions necessary for the manipulation of populations. These functions include the creation of populations and population members, the iterative population model, fitness scaling, parent selection and sampling, and the generation of population statistics. In addition, miscellaneous functions are included in the kernel (e.g., random number generators).

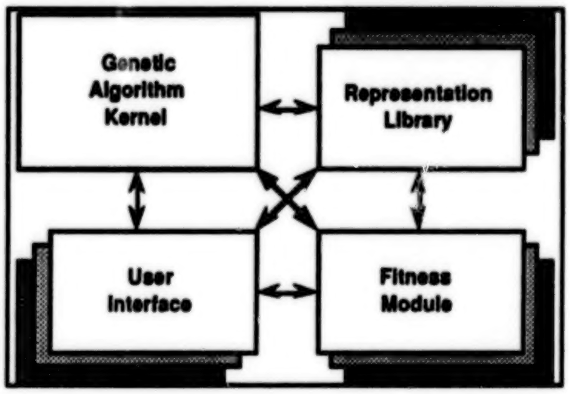


Figure 4. Splicer Project Design

Different types of problem-encoding schemes and functions are defined and stored in interchangeable *representation libraries*. This allows the GA kernel to be used for any representation scheme. At present, the Splicer tool provides representation libraries for binary strings and for permutations. These libraries contain functions for the definition, creation, and decoding of genetic strings, as well as multiple crossover and mutation operators. Furthermore, the Splicer tool defines the appropriate interfaces to allow users to create new representation libraries (e.g., for use with vectors or grammars).

Fitness functions are defined and stored in interchangeable *fitness modules*. Fitness modules are the only piece of the Splicer system a user will normally be required to create or alter to solve a particular problem. Within a fitness module, a user can create a fitness function, set the initial values for various Splicer control parameters (e.g., population size), create a function which graphically draws the best solutions as they are found, and provide descriptive information about the problem being solved. The tool comes with several example fitness modules.

The Splicer tool provides three *user interface libraries*: a Macintosh user interface, an X Window System user interface, and a simple, menu-driven, character-based user interface. The first two user interfaces are event-driven and provide graphic output using windows.

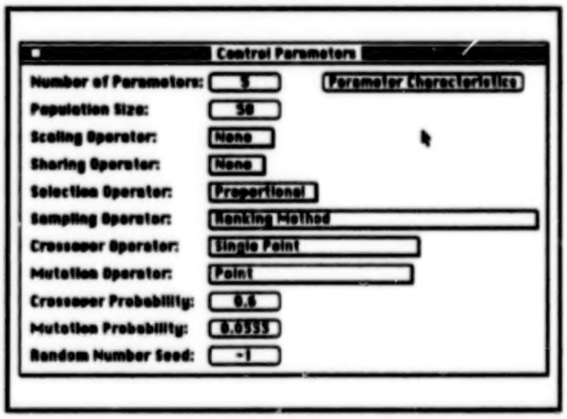
Implementation. The C programming language was chosen for portability and modularization. The original prototype was developed on an Apple Macintosh™ using Symantec's Think C™. This included the development of the Macintosh user interface. The GA kernel, representation libraries, and fitness functions were then ported to a Sun 3/80™ and SPARC™. An X Windows System user interface was then developed using X and the Hewlett-Packard Widget Set.

Results

This section will present the results to date of the Splicer Project. This will be done using brief descriptions and pictures of components from the Macintosh interface (components of the X Window System interface are very similar).

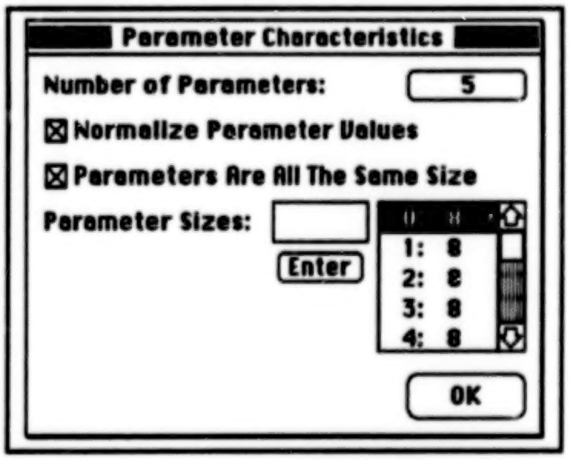
Control Parameters. The **Control Parameters** dialog box, shown in figure 5, allows the user to set the values of multiple parameters that control the operation of the Splicer tool (e.g., population size, crossover operator, mutation probability). This is accomplished in two ways: numeric parameters have buttons associated with them that pop up dialog boxes to allow the user to enter a new value; genetic operators (e.g., the fitness scaling operator) have pop up menus associated with them that allow the user to select from a list of operators.

The **Parameter Characteristics** button on the **Control Parameters** dialog box is used to pop up another dialog box that allows the characteristics of the individual problem definition parameters to be changed (for permutations there are no characteristics to change, therefore this button is disabled). The **Parameter Characteristics** dialog box for binary strings is shown in figure 6. This dialog box allows the user to specify the number of problem parameters and their size in number of bits. It also allows the parameter values to be normalized during decoding to create floating point numbers.



Number of Parameters: 5
Population Size: 50
Scaling Operator: None
Sharing Operator: None
Selection Operator: Proportional
Sampling Operator: Ranking Method
Crossover Operator: Single Point
Mutation Operator: Point
Crossover Probability: 0.6
Mutation Probability: 0.0333
Random Number Seed: -1

Figure 5. Control Parameters Dialog Box.



Number of Parameters: 5
☒ Normalize Parameter Values
☒ Parameters Are All The Same Size
Parameter Sizes:
0: 8
1: 8
2: 8
3: 8
4: 8

Figure 6. The Parameter Characteristics Dialog Box for Binary Strings.

Program Control. The operation of the Splicer program is controlled using options on the **Control** menu, shown in figure 7, as well as other menus (some of these are shown below). The **Control Parameters** dialog box is displayed using the **Set Control Parameters...** option. To create a random population, the **Create Population** option is selected. The **Run** option starts the execution of the genetic algorithms on the existing population. To begin again with a clean slate, the **Reinitialize...** option is used.

Operators. The Splicer program provides multiple genetic operators for fitness scaling, (in the future: fitness sharing), parent selection, parent sampling, crossover, and mutation. These operators can be changed at any time, even while the genetic algorithms are executing, and are selected using either the **Operators** menu (shown in figure 8) or the menu buttons on the **Control Parameters** dialog box.

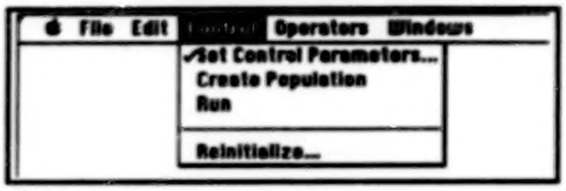

File
Edit
Control
Operators
Windows
✓Set Control Parameters...
Create Population
Run
Reinitialize...

Figure 7. The Control Menu.

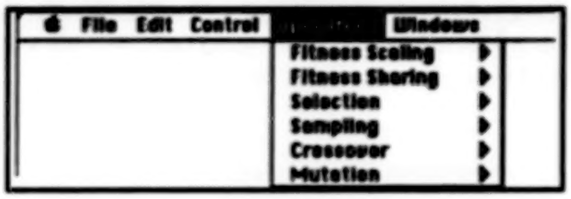

File
Edit
Control
Operators
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Crossover
Mutation

Figure 8. The Operators Menu.

Fitness Scaling. Splicer provides a linear fitness scaling option, as shown in figure 9. Fitness scaling is useful near the end of a genetic algorithm run when all members of the population have high fitness. Scaling spreads the fitness values and gives fitter members a higher reproductive advantage. Scaling can be turned off.

Fitness Sharing. While there is a place holder on the operators menu for fitness sharing, this option has not been implemented yet.

Parent Selection. Parent selection can be performed using either proportional selection (i.e., using relative fitness values) or using linear rank selection (where population members are simply ranked according to fitness), as shown in figure 10.

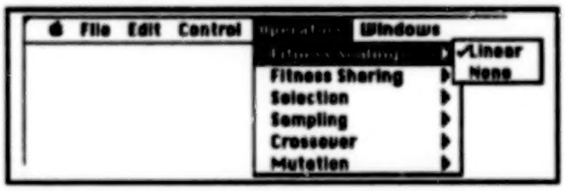

File
Edit
Control
Operators
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Crossover
Mutation
✓Linear
None

Figure 9. The Fitness Scaling Menu.

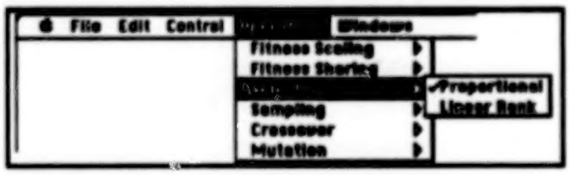

File
Edit
Control
Operators
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Crossover
Mutation
✓Proportional
Linear Rank

Figure 10. The Selection Menu.

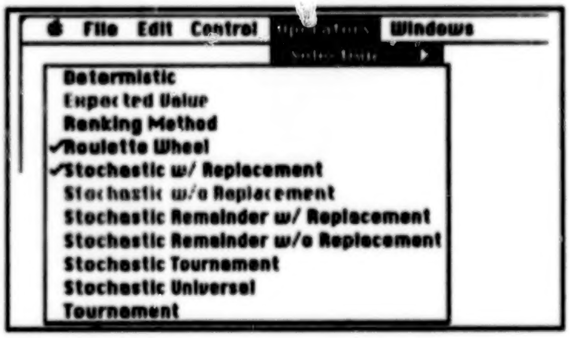

File
Edit
Control
Operators
Windows
Selection
Deterministic
Expected Value
Ranking Method
✓Roulette Wheel
✓Stochastic w/ Replacement
Stochastic w/o Replacement
Stochastic Remainder w/ Replacement
Stochastic Remainder w/o Replacement
Stochastic Tournament
Stochastic Universal
Tournament

Figure 11. The Sampling Menu.

Parent Sampling. Parents can be sampled using one of the methods shown in figure 11.

Crossover. The availability of crossover operators depends upon the representation library being used. The crossover operators provided by the binary string library are shown in figure 12. The crossover operators provided by the permutation library are shown in figure 13.

Mutation. Similarly, the availability of mutation operators depends upon the representation library being used. The mutation operators provided by the binary string library are shown in figure 14. The mutation operators provided by the permutation library are shown in figure 15.

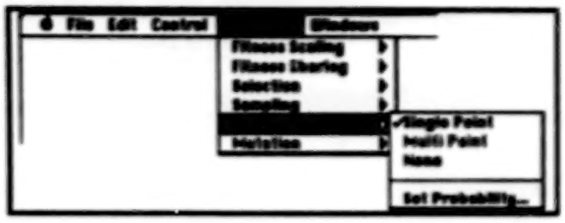

File
Edit
Control
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Mutation
✓Single Point
Multi Point
None
Set Probability...

Figure 12. The Crossover Menu Using the Binary String Library.

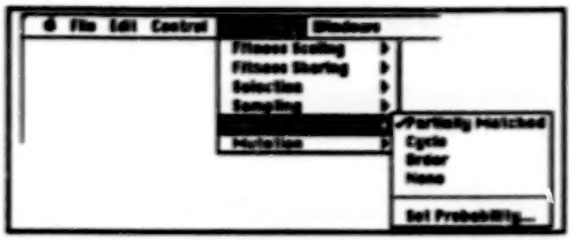

File
Edit
Control
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Mutation
✓Partially Matched
Cycle
Order
None
Set Probability...

Figure 13. The Crossover Menu Using the Permutation Library.

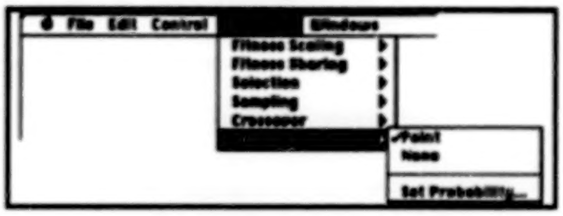

File
Edit
Control
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Crossover
✓Point
None
Set Probability...

Figure 14. The Mutation Menu Using the Binary String Library.

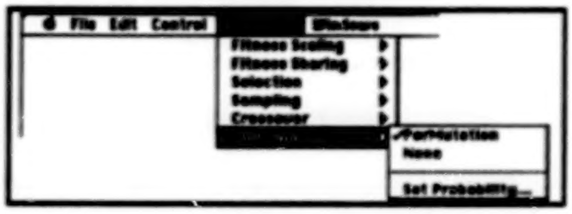
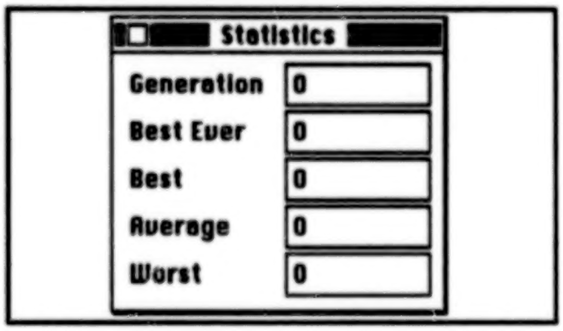

File
Edit
Control
Windows
Fitness Scaling
Fitness Sharing
Selection
Sampling
Crossover
✓Permutation
None
Set Probability...

Figure 15. The Mutation Menu Using the Permutation Library.

Output. Various windows present information to the user as the genetic algorithms execute. These windows are described briefly in the following sections.

The Statistics Window. The **Statistics Window**, shown in figure 16, displays the current generation number along with numeric objective function measures of the best solution ever found and the best, average, and worst members of the current population.

The Fitness Window. The **Fitness Window**, shown in figure 17, displays the fitness distribution of the current population. Fitness values are normalized, using the best ever fitness value, to create fitness values between 0.0 and 1.0. This interval is divided into 0.1 size bins and the percentage of the population in each bin is presented as a histogram.


Statistics

Generation	0
Best Ever	0
Best	0
Average	0
Worst	0

Figure 16. The Statistics Window.

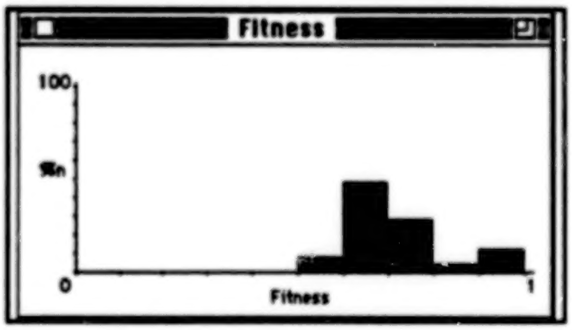

Fitness
100
%n
0
Fitness
1

Figure 17. The Fitness Window.

The Objective Window. The **Objective Window**, shown in figure 18, displays the numeric information from the **Statistics Window** but in historic form as a strip chart.

The User Window. The **User Window**, shown in figure 19, is controlled by the developer of the fitness module. The Splicer program provides simple drawing commands that the developer can use to draw in this window. The user is notified whenever a "best ever" solution has been found. One way this window can be used is to graphically display information about this solution. For example, in the window shown, the best

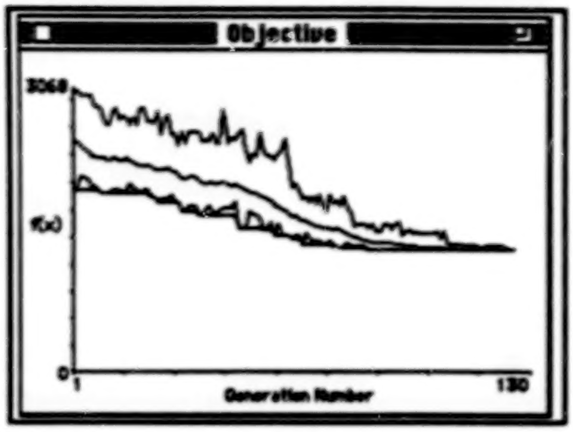


Figure 18. The Objective Window.

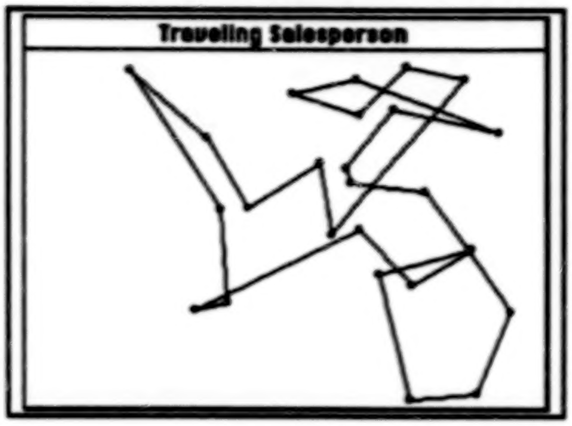
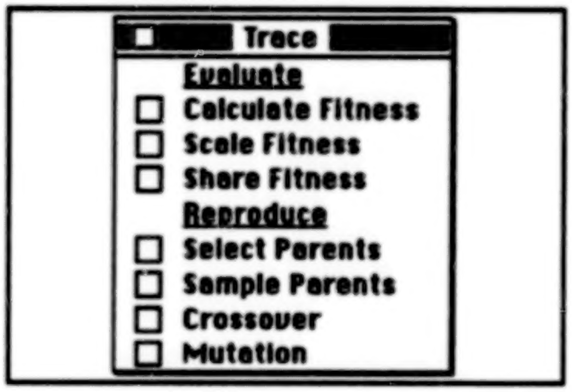


Figure 19. The User Window.

solution (so far) for a 25 city traveling salesperson problem is drawn. The numeric value of this solution would simultaneously appear in the **Statistics Window.**

The Trace Window. The **Trace Window**, shown in figure 20, presents the algorithm used by the genetic algorithms and highlights the current activity the genetic algorithms are performing by filling the appropriate box.



Evaluate
☐ Calculate Fitness
☐ Scale Fitness
☐ Share Fitness
Reproduce
☐ Select Parents
☐ Sample Parents
☐ Crossover
☐ Mutation

Figure 20. The Trace Window.

Status. An early version of the Macintosh-based Splicer tool was released for beta testing in July, 1990. During the beta testing period, the tool was ported to a Sun 3/80 and an X Window System user interface was developed. Significant new functionality, derived from the literature and suggestions from beta test feedback, and bug fixes were then incorporated into both versions. Documentation (user and reference manuals) is currently being developed. Version 1.0 of the Splicer genetic algorithm tool (both Macintosh and X Windows) will be released within the near future.

Future Work. Future work for this project includes the following:

- Additional functionality (e.g., *steady state* GAs, fitness sharing, other crossover or mutation operators)
- Genetic algorithm research
- Application of GAs and the Splicer tool within specific domains (e.g., job shop scheduling)
- Augmentation of the Splicer tool to create an application-specific tool within a particular domain

REFERENCES

1. Holland, John H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (Ann Arbor), 1975.

2. Goldberg, David E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wessley Publishing Company, Inc. (Reading), 1989.

3. Goldberg, David E.: Computer-aided Gas Pipeline Operation using Genetic Algorithms and Rule Learning. Doctoral Dissertation, University of Michigan, 1983.

4. Goldberg, David E.; and Samtani, M. P.: Engineering Optimization via Genetic Algorithms. Proceedings of the Ninth Conference on Electronic Computation, 1986, p. 471-482.

5. Fitzpatrick, J.M.; Greffenstette, J.J.; and Van Gucht, D.: Image registration by genetic search. Proceedings of IEEE Southeast Conference, 1984, pp. 460-464.

6. Baffes, P.; and Wang, L.: Mobile Transporter Path Planning Using a Genetic Algorithm Approach. Proceedings of SPIE's Cambridge Symposium on Advances in Intelligent Robotics Systems, Cambridge, MA, 1988.

N91

24051

UNCLAS

N91-24051

INTEGRATED VERTICAL BLOCH LINE (VBL) MEMORY

R. R. Katti, J. C. Wu, and H. L. Stadler
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109

ABSTRACT

Vertical Bloch Line (VBL) Memory is a recently conceived, integrated, solid-state, block-access, VLSI memory which offers the potential of 1 Gbit/cm^2 areal storage density, data rates of hundreds of megabits per second, and submillisecond average access times simultaneously at relatively low mass, volume, and power values when compared to alternative technologies. VBLs are micromagnetic structures within magnetic domain walls which can be manipulated using magnetic fields from integrated conductors. The presence or absence of VBL pairs are used to store binary information. At present, efforts are being directed at developing a single-chip memory using 25 Mbit/cm^2 technology in magnetic garnet material which integrates, at a single operating point, the writing, storage, reading, and amplification functions needed in a memory. This paper describes the current design architecture, functional elements, and supercomputer simulation results which are used to assist the design process.

INTRODUCTION

Vertical Bloch Line (VBL) Memory[1, 2, 3] is a solid-state, radiation-hard, nonvolatile, block access, magnetic VLSI memory. Research and development efforts for this novel memory are being pursued in the United States, Europe, and Japan. Table 1 shows the potential storage density that is achievable with VBL memory. The densities are a function of stripe width and line feature width, which are defined respectively by the magnetic garnet material and the lithographic process.

In a VBL memory, information is stored using VBL pairs in magnetic stripe domains in garnets. The presence or absence of a Vertical Bloch Line pair in a bit-cell location defines a binary "1"and "0," respectively. Input to the chip is performed by converting currents into magnetic bubbles and then into VBL pairs. Output sensing is performed by converting VBL pairs into magnetic bubbles and sensing magnetic bubbles magnetoresistively.

PRESENT DEVICE DESIGN

Fabrication

The present design uses the magnet garnet, $(BiYGdHoCa)_3(FeGeSi)_5O_{12}$, as the storage medium. The thickness, stripe width, collapse field, saturation magnetization, and anisotropy field of the film is approximately 2.4 μ m, 2.4 μ m, 230 Oe, 450 Oe, and 1800 Oe, respectively. The film is grown epitaxially on a non-magnetic gadolinium-gallium-garnet (GGG) substrate. These films are transparent but also have a large Faraday rotation, so that magnetic stripes, magnetic bubbles, and, under certain conditions, VBLs can be observed magneto-optically with polarized light using the Faraday effect in a polarized light microscope.

The magnetic garnet has perpendicular magnetic anisotropy so that the magnetization lies perpendicular to the film plane, with the bulk of the film magnetized in one direction, and the stripes magnetized in the opposite direction. A magnetic domain wall is the boundary between the stripe's magnetization and the magnetization of the rest of the film. A twist of magnetization in the domain wall in the plane of the film is a VBL, and two such twists form a VBL pair. If the chirality, or sense of rotation, of the VBLs in the wall is the same, the VBL pair is stable, with a size calculated to be much less than 1 μ m. The VBL pair is bound together energetically by VBL demagnetizing field energy and magnetic exchange energy.

C-2

The present device is divided into four main functional areas. First, the VBL storage area is designed to confine and stabilize stripe domains. Second, read/write gates are needed to convert VBLs to bubbles and vice versa. Third, a major line is needed for propagating bubbles which are used for input and output. Fourth, the output detector is needed for generating the output signal voltage. A sample architecture for a VBL chip is shown schematically in Figure 1.

The present device is built with ten mask layers as shown in Figure 2. Three metal mask layers are used for providing the contact pads and conductors which control the strip generator, read/write gates, bubble generator, and major line. SiO_2 and photoresist are used for insulating metal layers, and a window mask is used to open vias and contacts when needed between metal layers. Two ion implantation mask steps, using 150 keV Ne+ ions at a dosage of 4×10^{15} ions/cm^2, are used so that, after etching, grooves which are $0.2\,\mu$ m and $0.4\,\mu$ m thick are created. These grooves create stable locations in the garnet film for VBL stripe domains. An additional ion implantation mask layer is used in the major line to inhibit VBL formation in the input/output bubbles. A permalloy ($Ni_{0.8}Fe_{0.2}$) mask deposition is used to develop a magnetoresistive sensor which senses magnetic bubbles at the output and produces output voltages. A cobalt-alloy (i.e., CoPt) mask deposition is used for creating bit cells for VBL pairs along the walls of stripe domains. In test chips, a 5 nm thick Cr mirror layer is used to assist in stripe, bubble, and VBL observations using the magneto-optic Faraday effect during testing.

Supercomputer simulations[4, 5] which compute the effect of magnetic fields on magnetic domains, such as with stripe grooving, the major line, and the major line expander, are used heavily to assist in device design. VBL chip layouts are performed on an HP workstation and IBM PC/AT computers. Layouts are converted into fabrication masks using CIF and GDS II formats.

Stripe and Bit Stabilization

Data in the form of VBL pairs are stored in the domain walls of array of strip domains as depicted in Figure 3. The stripes are physically located in grooved regions in the garnet, as shown schematically in Figure 1 and in the design layout in Figure 4. The grooving allows selecting the bias field so that stripes are stable when the other chips functions, including the major line, are operating. The demagnetizing field produced by the bias field at the edge of grooved regions also serves to hold the stripe end and produce a stabilizing, effective edge-affinity magnetic field. Results from a supercomputer simulation of the formation of stable stripes in grooved garnet are shown in Figure 5. The computations were performed at 1 nsec time steps, and the computed stripe domain shape is shown at 40 nsec intervals.

Bit stabilization is used to stabilize VBL pairs along the stripe, as shown in Figures 1, 3, and 4. A periodic potential is placed along the stripe by an array of CoPt bars. For the CoPt, the saturation magnetization, coercivity, geometry, and spacing from the garnet are chosen to provide a sufficient field of approximately 5 Oe at the VBL stripe. This field value is currently considered to be enough to provide fields which create potential wells for the VBL pairs without disrupting the VBL pairs and moving stripes away from their groove-stabilized positions. The computed bit stabilization field profile at the end of an array with a $2\,\mu$ m bit period is shown in Figure 6. The distance between the CoPt and the garnet film is a parameter in the plot. The periodicity in the field profile is clearly evident.

Propagation of the VBL pairs, around the bit cells and to the read/write gates, can be achieved in two ways. First, a vertical pulse field can be applied which presses, or rocks, the stripe against the groove wall which gyrotropically causes VBL pairs to propagate down the stripe. Second, an in-plane field can be applied which directly causes VBL pairs to advance along the stripes' walls.

Read/Write Gates

Read/write Gates are used to convert VBL pairs in stripes into bubbles during the read process, and to convert bubbles into VBL pairs during the write process. It is necessary to read and write both "1's"and "0's"correctly. Nondestructive readback is achieved by using a current in conductors to bring the end of a stripe out of its groove into the read/write gate and into the presence of another conductor. If no VBL pair is present at the end of the stripe, the sense of the magnetization direction in the stripe wall causes the stripe to be difficult to chop with the field from a conductor because of the effect of exchange energy. The stripe is then returned to its stable position in the grooving. However, if a VBL pair is present, one VBL would be brought into the read/write gate while the other VBL would remain in the grooving. Hence, the chirality of the strip walls would be in the same direction which would readily allow the stripe to be chopped. The chopped portion of the stripe becomes a bubble, which is them propagated to the output for sensing, while the stripe returns to the grooved region. The chopping process recreates a VBL pair in the stripe which leaves the initial information intact.

Writing is achieved by bringing a bubble from the nucleator and major line to the desired read/write gate. If a bubble is present,m when the stripe is subjected to a field to bring it into the read/write gate, the stripe does not get drawn into the read/write gate because of magnetostatic repulsion between the bubble and the end of the stripe. Therefore, no writing to the stripe occurs. But if a bubble is not present, the stripe is brought rapidly into the read/write gate which inserts a VBL pair into the stripe, and the stripe is then allowed to relax into the grooving.

Major Line and Output Detector

The major line consists of a bubble nucleator for converting input signal currents into bubbles, a track for propagating bubbles from the nucleator to the read/write gates and from the read/write gates toward the output, an output detector for converting the demagnetizing field from bubbles into output voltages, and an expander which is used to stretch a bubble to a desired length to provide a satisfactory signal-to-noise ratio at the output.

A hairpin conductor is used for the nucleator. When a current is applied, such a conductor produces a magnetic field which is concentrated at the center of the hairpin. This field is used to generate bubbles which are used to transmit binary information to the VBL stripes via the read/write gates.

The propagation track in the major line consists of two levels of conductors. Each conductor is a sepentine arrangement of hairpin conductors which provide local magnetic field variations which form "waves"of stable positions for the bubbles down the track. The conductors are physically phase shifted by 90^0 to effect propagation. The principle of operation of the major line is shown in Figure 7. The layout of the major line track, along with the bubble nucleator and two read/write gates and grooves, is shown in Figure 4.

The output detector consists of a rectangular strip of permalloy, which is magnetoresistive. When the fringing magnetic field from a magnetic bubble affects the sensor, the sensor's resistance changes. When a reference current is issued to the sensor, the presence or absence of a bubble induces two different voltage levels which define binary "1's"and "0's." If it were necessary to maximize common-mode rejection from stray magnetic fields, two magnetoresistive sensors, including the actual sensor and a dummy sensor, can be used which allow signals to be measured differentially. Signals from the magnetoresistive sensor are increased if longer elements with greater electrical resistance are used. Bubbles can be stretched in length, to provide additional magnetic fields for the lengthened sensor, by widening the bubble track as the bubble approaches the sensor. The combined output sensor and expanding major line are shown in Figure 8, in which only one propagation metal layer in the major line is shown for clarity. Shown in Figure 9 is a supercomputer simulation, shown at 40 nsec intervals after being computed at 1 nsec intervals, of a bubble being stretched and then unstretched as it passes through the expander and detector.

EXPERIMENTALRESULTS

Sample experimental results of chip functions from past designs are now presented.[4,7] These data were taken on test chips using a 4.76μ m thick $(BiYSmLu)_3(FeGa)_5O_{12}$ garnet film with a saturation magnetization of 200 Oe, zero-field stripe width of 4.67 μ m, characteristic length of 0.61 μ m, domain wall mobility of 350 cm/sec/Oe, anisotropy field of 1350 Oe, coercivity of 1.1 Oe, and collapse field of 100 Oe.

Figures 10 and 11, respectively, show stabilized stripes in grooves near read/write gates at the end of a groove, and near the center of the stripe groove near the stripe nucleator. The stripes are observed, with polarized light using the magneto-optic Faraday effect, as white strips against a darker background. The dark bars running perpendicular to the stripes are the cobalt bit-stabilization bars as described in Figures 1, 3, and 4.

Figure 12 shows nucleated bubbles, as white spots, in the major line as they propagate down the track. Bubbles and stretched bubbles in the expander portion of the major line near the output detector are shown as white sports and strips, respectively, in Figure 13.

CONCLUSION

VLSI designs, simulation results, and experimental results have been presented which describe current work on the storage and input/output functions for solid state, high density, nonvolatile, radiation hard, block access Vertical Bloch Line (VBL) Memory. Such a memory offers the potential of achieving 1 Gbit/cm^2. Three metallizations are used along with three ion implantation steps, one permalloy magnetoresistive sensor deposition, and one CoPt bit-stabilization deposition. Previous and current experimental results and supercomputer simulations indicate that individual storage, read/write gate, and input/output functions are feasible. Present work is aimed at integrating all necessary memory functions on a single chip to achieve simultaneous operation, a unique operating point, and a fully-functional single chip memory.

ACKNOWLEDGEMENTS

The research described in this paper was performed by the Center for Space Microelectronics Technology, Jet Propulsion Laboratory, California Institute of Technology, and was sponsored in part by the Strategic Defense Initiative Organization Innovative Science and Technology Office. The use of the JPL/Caltech CRAY X_MP/18 for performing portions of this work is acknowledged.

REFERENCES

1. S. Konishi. "A New Ultra-High-Density Solid State Memory: Bloch Line Memory," IEEE Transactions on Magnetics, Vol. 19, No. 5, pp. 1838-1840, September, 1983.

2. T. Suzuki, et al. "Chip Organization of Bloch Line Memory," IEEE Transactions on Magnetics, Vol. 22, No. 5, pp. 784-789, September, 1986.

3. H. Matsutera, et al. "Design and Characteristics for Vertical Bloch Line Memory Using Ring-Shaped Domain," IEEE Transactions on Magnetics, Vol. 23, No. 5, pp. 2320-2325, September, 1987.

4. J. C. Wu. "Solid-State Magnetic Devices: Magnetic Bubble Logic Devices and Vertical Bloch Line Memory," Ph.D. Thesis, Carnegie Mellon University, August, 1986.

5. N. Hayashi and K. Abe. "Computer Simulation of Magnetic Bubble Domain Wall Motion," Japanese Journal of Applied Physics, Vol. 15, No. 9, pp. 1683-1694, September, 1976.

6. J. C. Wu, et al. "Major Line Operation in Vertical Bloch Line Memory," Paper FP-10, Accepted for Presentation, 1990 Magnetism and Magnetic Materials Conference, San Diego, CA, October, 1990.

7. J. C. Wu, et al. "Stripe Stabilization in Vertical Bloch Line Memory," Paper FP-9, Accepted for Presentation, 1990 Magnetism and Magnetic Materials Conference, San Diego, CA, October, 1990.

	$L_f = 1\,\mu m$	$L_f = 0.5\,\mu m$	$L_f = 0.1\,\mu m$
$S_w = 5\,\mu m$:	10 Mbits/cm^2	20 Mbits/cm^2	100 Mbits/cm^2
$S_w = 2\,\mu m$:	25 Mbits/cm^2	50 Mbits/cm^2	250 Mbits/cm^2
$S_w = 1\,\mu m$:	50 Mbits/cm^2	100 Mbits/cm^2	500 Mbits/cm^2
$S_w = 0.5\,\mu m$:	100 Mbits/cm^2	200 Mbits/cm^2	1000 Mbits/cm^2
$S_w = 0.25\,\mu m$:	200 Mbits/cm^2	400 Mbits/cm^2	2000 Mbits/cm^2

Table 1: VBL Storage Density as a Function of Stripe Width, S_w, and Line Feature Size, L_f. Storage density is inversely proportional to $2S_wL_f$.

Counter-clockwise stripe without static in-plane field

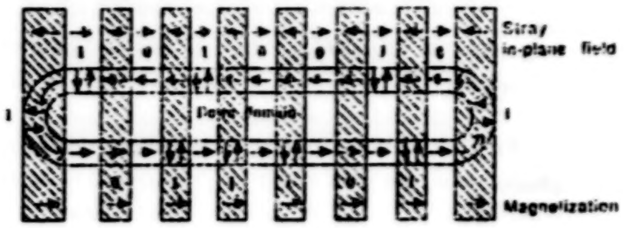

Stray in-plane field
I
Down domain
I
Magnetization

Figure 3: Schematic of VBL Stripe Storage Structure.

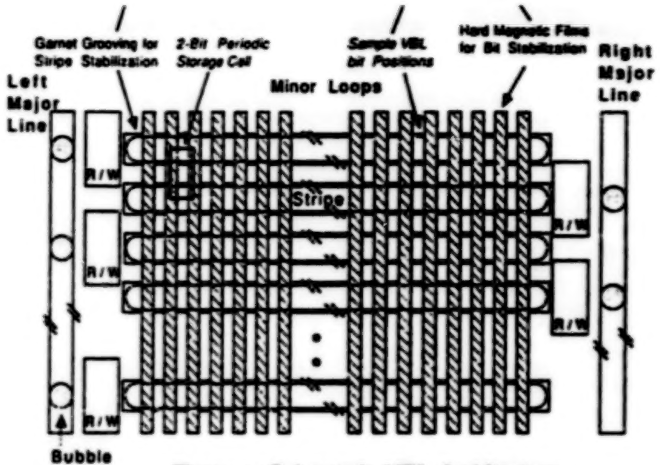

Garnet Grooving for Stripe Stabilization
2-Bit Periodic Storage Cell
Sample VBL bit Positions
Hard Magnetic Films for Bit Stabilization
Left Major Line
Minor Loops
Right Major Line
R/W
Stripe
Bubble

Figure 1: Schematic VBL Architecture.

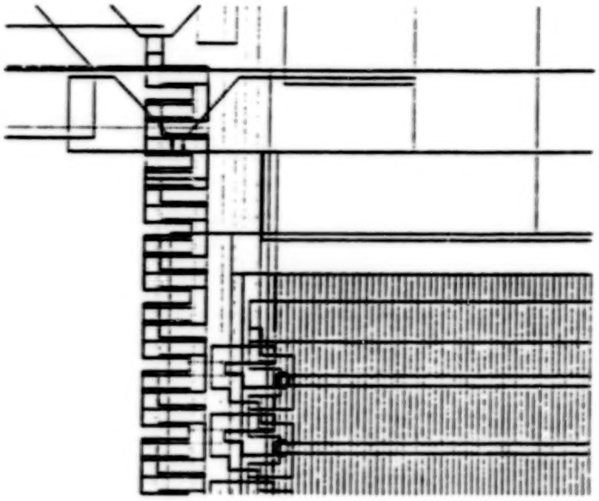


Figure 4: Layout of VBL Chip Nucleator, Dual-Conductor Major Line, and T Read/Write gates and Two Stripe Grooves.

Cut-away View of VBL Memory Devices

Process Option

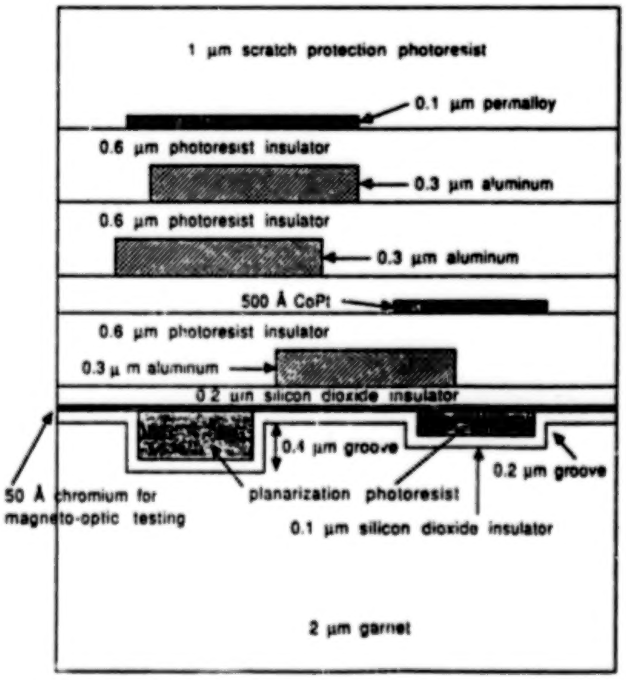

1 μm scratch protection photoresist
0.1 μm permalloy
0.6 μm photoresist insulator
0.3 μm aluminum
0.6 μm photoresist insulator
0.3 μm aluminum
500 Å CoPt
0.6 μm photoresist insulator
0.3 μ m aluminum
0.2 μm silicon dioxide insulator
0.4 μm groove
0.2 μm groove
50 Å chromium for magneto-optic testing
planarization photoresist
0.1 μm silicon dioxide insulator
2 μm garnet

Figure 2: Cross-Section of the Current VBL Fabrication Process.

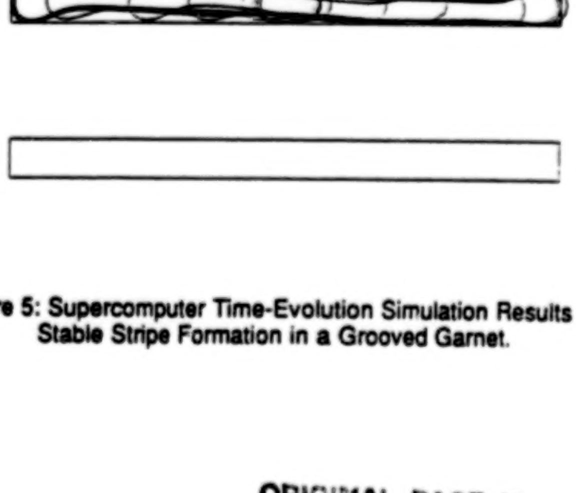
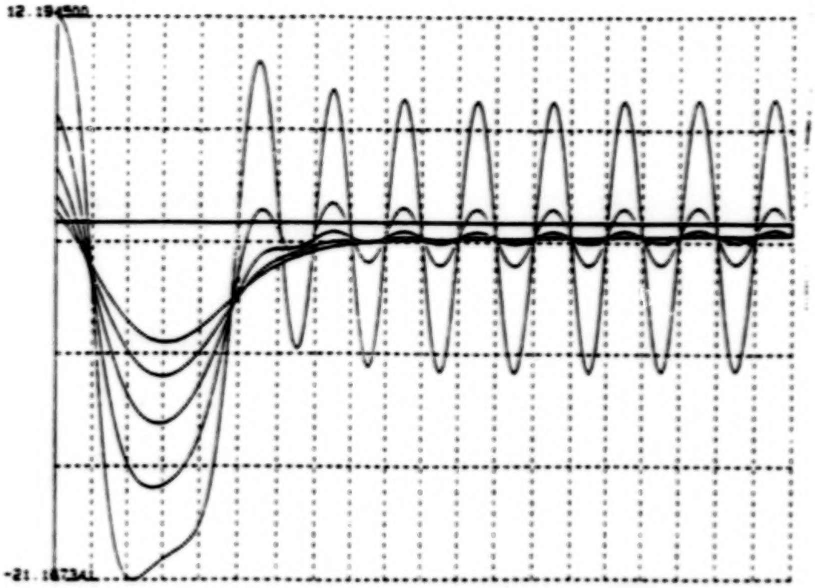


Figure 5: Supercomputer Time-Evolution Simulation Results of Stable Stripe Formation in a Grooved Garnet.

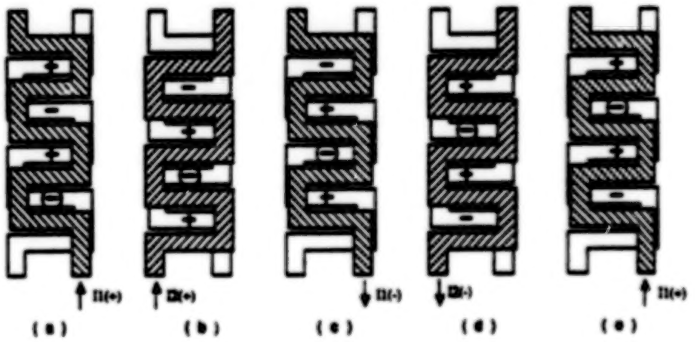
ORIGINAL PAGE IS
OF POOR QUALITY



```
The parallel bars are infinitely long in the y-axis  Field max=12.194500, min=-21.187341
a1=10.0, a2=11.0, c1=0.00, c2=0.04, sp= 1.0, 4piMs= 5000 G, xa= 4.000, xc=24.000
za= 1.040, zc= 3.040, gridx=10, gridz= 4, NB=100, in-plane component
```

Figure 6: Computed Periodic Magnetic Field Profile of VBL Bit-Stabilization Wells.

Propagation of Bubbles Using Current



A current in a zig-zag conductor, as shown in (a), produces an alternating magnetic field pattern which attracts or repels bubbles to move along the conductor. By staggering two conductor patterns as shown, and
alternating the current sent to each conductor, a track is defined which can propagate bubbles.
Due to the low inductance of the conductor sheet, such a propagator can be operated at a very high frequency.

Figure 2.4.1.: Principle of Bubble Propagation in a Dual-Conductor Major Line.

Figure 7: Principle of Bubble Propagation in a Dual-Conductor Major Line.

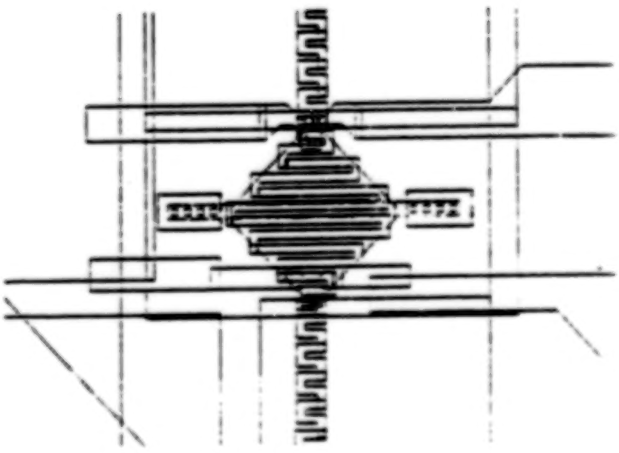


Figure 8: Layout of VBL Expander and Output Detector.

ORIGINAL PAGE IS
OF POOR QUALITY

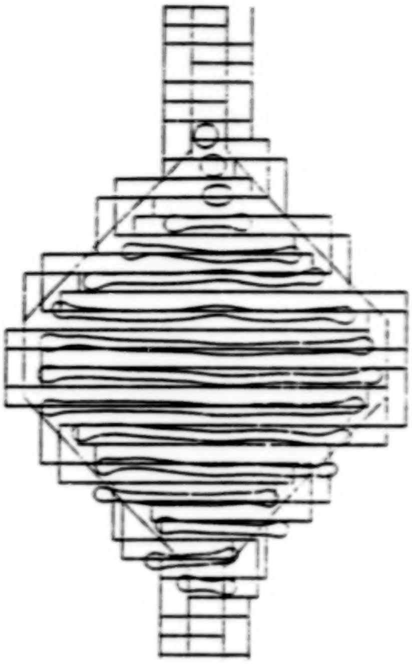


Figure 9: Supercomputer Time-Evolution Simulation of an Expanding Bubble in the Major Line Expander and Output Detector.

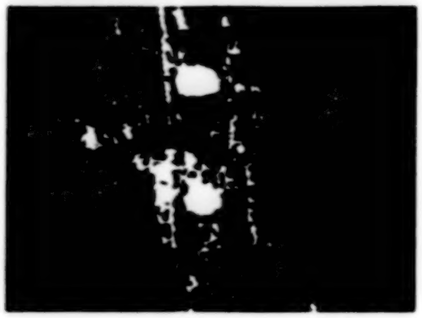


Figure 12: Major Line Under Operation Near the Center of the Major Track.

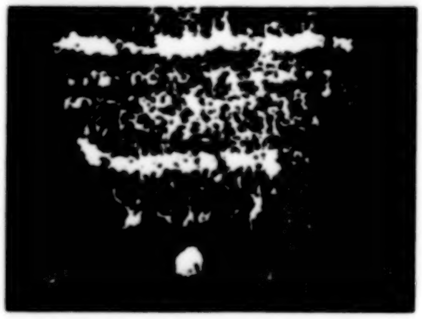


Figure 13: Major Line and the Major Line Expander Under Operation.

Figure 10: Photograph of Stable Stripes Near the Read/Write Gate end in a VBL Test Chip.

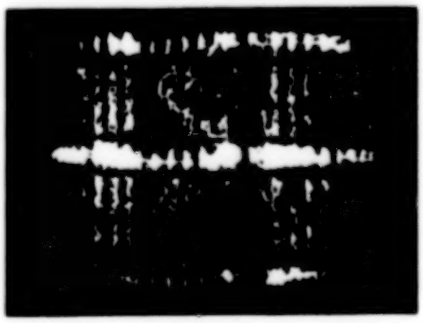


Figure 11: Photograph of Stable Stripes at the Stripe's Center, Near the Stripe Nucleator in a VBL Test Chip.

ORIGINAL PAGE IS
OF POOR QUALITY

SESSION C - ENVIRONMENTAL TECHNOLOGY

Wednesday November 28, 1990

- **Physical/Chemical Closed-Loop Water Recycling**
- **Water Quality Analyzer**
- **New Research On Bioregenerative Air/Water Purification Systems**
- **Environmental And Facilities Management System (EMFS)**
- **The Land Analysis System**
- **Potential Commercial Use Of EOS Remote Sensing Products**

N91

24052

UNCLAS

N91-24052

Physical/Chemical Closed-loop Water-recycling

Cal C. Herrmann Bionetics Inc., NASA Ames Research Center, MS 239-11, Moffett Field, CA 94035

Theodore Wydeven NASA Ames Research Center, MS 239-4, Moffett Field, CA 94035

ABSTRACT

Water needs, water sources, and means for recycling water are examined in terms appropriate to the water-quality requirements of a small crew and spacecraft intended for long-duration exploration missions. Inorganic, organic, and biological hazards are estimated for wastewater sources. Sensitivities to these hazards for human uses are estimated.

The water recycling processes considered are humidity condensation, carbon dioxide reduction, waste oxidation, distillation, reverse osmosis, pervaporation, electrodialysis, ion exchange, carbon sorption, and electrochemical oxidation. Limitations and applications of these processes are evaluated in terms of water-quality objectives. Computer simulation of some of these chemical processes is examined.

Recommendations are made for development of new water recycling technology and improvement of existing technology for near-term application to life support systems for humans in space. The technological developments are equally applicable to water needs on earth, in regions where extensive water-recycling is needed or where advanced water treatment is essential to meet EPA health standards.

INTRODUCTION

This report addresses minimum water quality requirements in terms of space mission needs. Means to achieve, assure and exceed that water quality are discussed. The fundamental approach is to adapt and extend established desalination and water-treatment technology to the special requirements of: (1) a high degree of water recovery (90% and more), (2) a small habitat, and (3) energy-efficiency.

A spacecraft, in a substantially closed-loop mode, may be treated as a microcosm of the open environment of Earth: the sources and uses of water are discrete, the pathways definable, and the reservoirs limited. As a closed system, a mass balance must be achieved, both for the water and for its solutes. The sources of used water for recycling will be identified first, then the water needs. The water-cycle options are thus limited, and a set of requirements for water quality and treatment is implied.

Desalination and other means of achieving the water quality required are then examined, with concern about technical and practical limitations of the methods and their appropriate applications in spacecraft. The recommendations for treatment yield another set of implications for water quality and implicate monitoring requirements to assure that the water remains satisfactory and to indicate repair or replenishment actions.

Susceptibility to hazards is rated by category:

1) Inorganic (e.g. corrosion contaminants),
2) Organic (e.g. detergents), and
3) Biological (e.g. viruses and microbes).

It is always preferable to avoid hazardous water contaminants (beyond normal uses), but this is not possible when contamination occurs:

1) from normal aging processes of the spacecraft: outgassing of plastics and normal corrosion, or

2) from accidental processes such as overheating of components, resulting in degradation of plastics and enhanced corrosion of metals, or

3) as residual viruses or other pathogens, including possible mutation of previously tolerated viruses during the flight, and the remote possibility of pathogens introduced from Mars.

One or more examples of performance of water purification systems can be created, based on reasonable choices of the water cycle, with the flows defined and the streams analyzed first by spreadsheet calculation of mass balances around the loop(s) for water and the major solutes and then by computer simulation (1) of the water treatment processes. Computer simulation allows rapid testing of the system's response under normal and abnormal operating conditions, and the simulation itself can be tested for validity by a few laboratory experiments.

WATER SUBSYSTEM OPTIONS

Most of the sources and needs are the same as, or similar to, those projected for Space Station Freedom. However, for the longer Mars mission the transport cost of an accumulated shortfall (needs exceeding the recycled water) is greater.

Recyclable water sources

Condensate - Condensate from the humidity-control condenser originates from human breath and perspiration plus evaporation from other water: showers, clothes-washing, etc. This source can hardly be expected to be free of viruses or bacteria of concern to humans. Though the crew in the closed environment clearly already have been exposed to air-borne infection, microbes could get out of control growing outside the body's active protections. An amount of corrosion contaminants and products of plastic (or lubricant) degradation is possible from the equipment in the ship.

Carbon dioxide reduction - A result of reduction of respired carbon dioxide by the Bosch or Sabatier processes is 1.15 to 1.8 pounds/person-day of water (2). This water may be added to the potable stream, since its source streams (CO_2 and H_2) are gas phases of purity at least comparable to cabin condensate, and is heat-sterilized — although ion-exchange protection against corrosion products is advisable.

Wash water - This waste water is from laundry-, shower-, hand wash-, and dish-water. It is clearly lower in quality than the condensate water. Measurable quantities of microbial, organic, and inorganic contaminants have been found and would be accumulated if not removed before recycling. Corrosion products from normal skin contact with equipment and tools are in this water.

Laundry water is the largest hygiene water element. It can be treated separately from the water used for washing people if different quality standards are applied for laundry use than for human contact. For the purpose of tracking constituents and testing input water assignments, it will be separated on a spreadsheet through the processing.

Urine - Studies for Space Station Freedom have demonstrated that potable water can be made from urine, using distillation and sorption. However, if adequate water of potable quality is available from a higher-quality source (lower solute and bacteria contents), there is no need to use recycled water from urine for other than laundry or electrolysis.

Waste oxidation - Oxidation of solid wastes, including wrapping materials, paper, tissues, fecal material, and hygiene aids, is a subject of another study parallel to this one. The useful result of the process of waste oxidation and reduction of formed carbon dioxide to carbon, of a waste of average composition CH_2O, is water. The amount is significant. From an estimated total of 2.2 pounds of waste or trash/person-day, more than one pound per person-day of new water is predicted. This provides an increasing reserve of greater than 2% of a person's requirements per day, potentially compensating for accidental or functional losses. This accumulation also offers the possibility that water needed by the Mars landing crew need not be returned to the spacecraft, conserving lift-off power.

Sources	Hazard Estimates		
	Inorganic	Organic	Biological
Condensate	slight	slight if filtered air	significant
CO2 Reduction	slight	very small	slight
Wash Water	significant	significant	significant
Waste Oxidation	corrosion possible	slight	slight
Urine	substantial (NH3+salts)	substantial	substantial

Water needs

Potable. - Potable water is needed for drinking, for food preparation (hydration and washing), for hygienic and medical uses where internal contact is expected (eye wash, brushing of teeth, occasional mouthwash and douches), and for possible other uses such as final rinsing of dishes.

Laundry - The largest single use of water is for washing clothes, over half of the "hygiene water" requirement. In an evaluation of U.S. Navy shipboard laundry-water recycling, moderate accumulation of organics and salts was tolerable through a recycling system containing flocculation, filtration, and activated carbon elements (3). A total organic carbon level of about 140 ppM was a steady-state condition. One would initially conclude that the water-quality standard for clothes-washing water may be substantially lower than that for potable water so the implications of partial purification are worth examination. These are: (1) Possible increase in use of carbon adsorbent. Since more organic material was allowed in the laundry-recycle water, passage through carbon adsorbent would be expected to remove more material and more rapidly exhaust the adsorbent, and (2) Possible increased use of bactericide. Increased organic content may consume more iodine or chlorine, also yielding more halocarbon residue. The Navy study did not use a bactericidal oxidant for their laundry water recycling scheme, but did use a bleach in the wash. A medical need for sterility of laundry water is not obvious. The only obvious access to the body of such residues is the use of washed handkerchiefs and dishtowels.

Significant water is evaporated, and shows up as condensate water. Lard et al. (3) find this as 5.6% (24 of 425 gallons) of the total laundry wash water, or about 1/4 of the final rinse water. A plausible target range for total dissolved solids (TDS) is 500-1000 ppM.

Hygiene wash - The second largest water use is washing persons and dishes. This water must be substantially pathogen- and toxin-free. Excess iodine or other bactericide can be tolerated, since the amount ingested by accident is surely small. Total dissolved solids (TDS) requirements are mild, assuming that adequate washing action is achieved. There is an advantage in maintaining at least marginal potability, since this water could serve as a backup in case of failure in the potable water supply. Here also water is lost to evaporation.

Other - Other water requirements are not well defined. If urinal flush water is only used to clean the collection device, filtered used shower water could be satisfactory, or even superior if residual detergent and bleach are present. Water for electrolysis, to produce hydrogen for reduction of respired carbon dioxide, has different quality requirements, since the presence of ions improves conductivity, reducing power losses, but some contaminants (e.g. sulfides) may poison catalytic electrodes. The first Working Group for the Space Station Freedom identified 0.2 to 0.4 pounds (mass, lbm) of hydrogen per person-day as required for the Bosch or Sabatier process for reduction of respired carbon dioxide (4). Some water is normally used for cleaning work surfaces. Water of many different levels of purity requirements may be required for experiments and process development space tasks.

Uses	Hazard Sensitivities		
	Inorganic	Organic	Biological
Potable	significant	significant	substantial
Laundry	slight	slight	slight
Hygiene-wash	slight	slight	slight

Implied water-quality requirements

The water quality required for these needs can be individually analyzed for each need. The lowest quality requirements are for laundry water, where controls of pH, oxidation potential, conductivity, turbidity and total dissolved carbon are sufficient for satisfactory performance.

For bodily-contact hygiene water, higher standards (lower concentrations) are appropriate for these measured quantities. Standards for potable water are well established (JSC-SPEC-SD-W-0020, NASA STD-30000 and EPA 1986). Optimal compositions of desalinated water based on palatability as well as World Health Organization (WHO) standards have been proposed by Gabrielli (5,6).

An estimate of the principal air contaminants expected in a space station has been made by Yoshimura et al. (7) and is used later in this report to estimate organics condensing with water in the humidity control system. A more recent similar table was published by Leban and Wagner (8). For a selection of those of greatest concentration, threshold limit values (TLV) values from Verschueren (9) can be converted into allowed daily dosage (based on an 8-hour day, 5-day work week), and converted into the equivalent exposure in drinking water. These values are considerably higher than EPA maximum allowable concentrations (MAC) for priority pollutants for the general population. Some explanations are: (1) occupational limits may be at levels of the beginning of impairment, (2) the exposures are substantial for a limited time, (3) and for a small part of the population assumed in good health, (4) none of the major constituents are in the critical first Safe Drinking Water Act set.

In practice, one can distinguish and apply three independent criteria for water quality:

1) Toxic: exceeding this limit can incur performance degradation and short or long term health effects.

2) Risk: principally cancer. A risk "limit" depends on the risk level accepted. The selection of an acceptable risk level is influenced by:

a) the size of the population subject to risk,
b) comparison with other risk levels accepted, and
c) the personal standards of those individuals subject to risk.

3) Organoleptic: unpleasant or excessive taste or odor. Organoleptic levels may be temporarily exceeded without harm, but performance may suffer in extreme situations. Nausea may be exacerbated by an unpleasant taste or odor. This factor could be significant for hygiene and laundry waters insofar as odors are released to the air.

Probably the best poll of opinion of individuals subject to risk would be of former and present astronauts, as an estimate of the risk standards of future astronauts. One might ask if a risk level for exposure to cancer hazards in air and water might reasonably be 1% of the accepted cancer hazard from radiation, or some other value. For a large population, such as an eventual space colony, this level of risk could be lowered.

RECLAMATION AND WATER QUALITY

Distillation

A vapor-compression distillation (VCD) device has been evaluated for Space Station Freedom use (10). Some disadvantages of this distillation technique have been found, but its advantages are high water recovery, potentially good isolation from solute and biological contaminants, and ease of mechanical repairability.

Development at the Water Technology Center of the University of California, Berkeley, for the State of California Department of Water Resources resulted in substantial improvement in the overall heat-transfer coefficient between the condensing and evaporating films: this has considerably reduced the energy requirement of VCD. Designing the system as a multistage unit greatly increased the water output for a given heat input and aided in reaching a high concentration of brine in a continuous, rather than batch, process (11, 12). A multistage configuration uses the heat from condensing the first distillate to evaporate a second amount of wastewater: the temperature and pressure drops per stage are additive. Increasing the heat transfer coefficient by thinning the condensate water films decreases the temperature drop per stage, allowing more stages for a given temperature difference between the feed and the final condenser.

Reverse osmosis

Recovery - Reverse osmosis (RO) is a hyperfiltration process involving high pressures (100 to 1000 psi) to press water through membranes designed to reject ionic and other larger species. It has been demonstrated to be appropriate for preparation of water for hygiene use and satisfactory as a source of potable water. Extraction of high-quality water from wash water containing of the order of 1000 ppM of total dissolved solids (13) can yield a 90% recovery, with a brine remaining of approximately 10 times that concentration. The brine could then pass to another process, such as oxidation of organics and/or distillation. To attempt a much higher recovery than 90% from such wash water presents the same problem as applying RO to water recovery from urine: higher pressures are required to overcome the osmotic pressure of a concentrated solution.

Permeability - Qualification of RO for extended use should include testing for extended periods with simulated or real washwater, since the plastics commonly used as membrane materials do have measurable permeabilities to organic solutes (14). The permeability of polymers to volatile organics can be turned to an advantage. A membrane selected on the basis of low porosity but appropriate solubility parameters, and thus impermeable to water, may allow permeation of organics in wastewater. In effect, this replaces the "air stripping" used in municipal water treatment but not possible in a confined atmosphere. Haxo et al. (15) give the following vapor transmission rates (VTR) of organic solvents and water passing through a 0.85 mm thickness of high-density polyethylene. Comparable data are also given for many other polymeric membrane materials.

Solvent	VTR ($gm/m^2/day$)	From partial pressure (mm Hg) (23°C):
Water	0.0472	10.53
Methanol	0.50	112
Acetone	2.19	212
Cyclohexane	151.	89
Xylene	212.	7
Chloroform	506.	178

Table 1. Vapor transmission rates through high-density polyethylene membranes. (15)

Electrodialysis

As a process that is electrically controllable and that operates at ambient temperature and pressure, electrodialysis (ED) is of interest for small desalination systems. It may be thought of as a self-regenerating ion-exchange-resin system, effective for removing ions but not organic solutes. Electrodialysis has been shown useful for further concentrating RO brines (16). At brine concentrations over 1%, the osmotic pressure to be overcome by pumping pressure becomes substantial in RO, while in the same concentration range for electrodialysis the increasing solution conductivity substantially reduces electrical power loss.

Ion exchange

A traditional use of ion exchange in water treatment has been for water softening, not needed in most water recycling. However the water softening process works by means of a useful effect: in dilute solutions polyvalent ions (calcium or magnesium in water softening) bind to a cation-exchange resin much more strongly than do monovalent ions: thus the resin releases sodium in exchange for calcium. Most of the potentially toxic heavy-metal ions of concern here are polyvalent (e.g. Cr^{+3}), and so are even more strongly bound than calcium. By careful choice from the many resins commercially available, including chelating resins, a guard column or cartridge can be tailored to the needs of the mission. Prediction of resin-water equilibria in the presence of several ionic species can be performed by the method of Klein (17); for use on small computers the Klein program has been revised into BASIC (18).

Carbon adsorption

Removing organic constituents, even at trace levels, is accomplished with activated carbon. Regeneration is a thermal process: desorbed gases must be vented or, preferably, oxidized. Oxidation could be performed if the adsorbates could be desorbed into an air stream supplying a solid- or liquid-waste oxidation processor. Removal capacities for many compounds are in the range of 100 mg per gram of carbon (19), but some volatile compounds of importance (e. g. methanol) are poorly adsorbed, and are retained only at loadings of the order of 10 mg per gram of carbon, (20: Table 17.2). An estimate of the amounts of organics to be removed from the water streams is necessary in order to choose the size of activated-carbon cartridges and to determine the need for and frequency of regeneration.

Other

Reports from development of supercritical water oxidation equipment for waste disposal (21) indicate that this may be a substantial source of water for recycling. Product sterility is certainly assured. Inorganic content is predicted to be low, from the insolubility of ionic compounds in the supercritical phase. A recent publication on this project indicates that further development is needed (22).

A convenience for future spacecraft missions incorporating food production using higher plants is that the supercritical oxidation process with minor modification can yield a nutrient water for growing plants, with nitrogen retained as ammonia (23) and inorganic nutrients redissolved on removal from the supercritical condition.

At first glance, an electrochemical oxidation proposed (24) does not seem competitive with the wet oxidation method, in that power is consumed in generating the oxidant electrochemically (25: p. 42). However, as much or more power may be required to make the oxygen consumed by any other oxidation process.

Bactericides

Bacteria can thrive in most water, so safety can never be assumed. Bacterial growth should be prevented even in the humidity condenser, which is poorly accessible to additions of disinfectants. One approach might be to add a volatile oxidizing agent as bactericide at the entrance to the condenser.

Sterility of final-product potable water must be separately assured. Concern has been expressed (26, 27) about the accumulation of iodide in water treated with enough iodine to maintain bactericidal activity during storage. Since iodine is readily formed by electrochemical oxidation of iodide, electrochemically recycling the reduced iodide to iodine would seem possible. In the low conductivity of potable water this would not be a rapid reaction, but if stored water is slowly circulated past the oxidizing electrode, disinfectant activity may be maintained with a small dosage of iodine.

Competitive oxidants can be prepared electrochemically as required: chlorine, hydrogen peroxide and/or ozone. The destruction of viruses by ozone is more rapid than by chlorine (28), which may be advantageous in the small water treatment module. Better taste and lack of halocarbon formation are additional advantages. Gnann (29) has found optimum ozone production to require an electrochemical cell with a concentrated phosphate buffer containing fluoride and cooling.

PERFORMANCE CHOICES

A first example is a simple configuration for water recycling. A second example indicates how process options can be applied. Such examples provide an opportunity to examine the means for trapping identified hazards and to search for failure modes.

Example 1

Flow choices - One configuration satisfying the criteria is shown in Figure 1. Humid air from the crew's respiration and from evaporation of washwater arrives at the condenser. Product water from CO_2 reduction is of high quality and can add to condensate water or to hygiene-wash, as needed. Initially it will be assumed that a particulate filter is not required, but that a carbon filter is, probably lightly charged with iodine. A cation-exchange resin is included to trap most inadvertent metals. A few ppM of calcium bicarbonate are added to improve taste, lower corrosiveness, and buffer pH (5, 6).

Washing-quality hygiene water comes from wash water recycled through RO or distillation. A carbon filter should be sufficient protection for this stream: very little bodily exposure to hazardous ions exists. Electrochemical chlorination (chloride in the RO brine, originating from sweat, is oxidized to chlorine) or ozonation might be more convenient than iodination, and appropriate if the source of this water is more subject to bacterial contamination. A field-shower recycling study (30) implied that high-quality water is not needed here, but the advantage of having an emergency source of potable water, combined with the potential inconvenience of transmission of disease or toxins through inadvertent access to the body led to the recommendation of RO or distillation purification here, with simpler treatment being applied to the more major water use that follows.

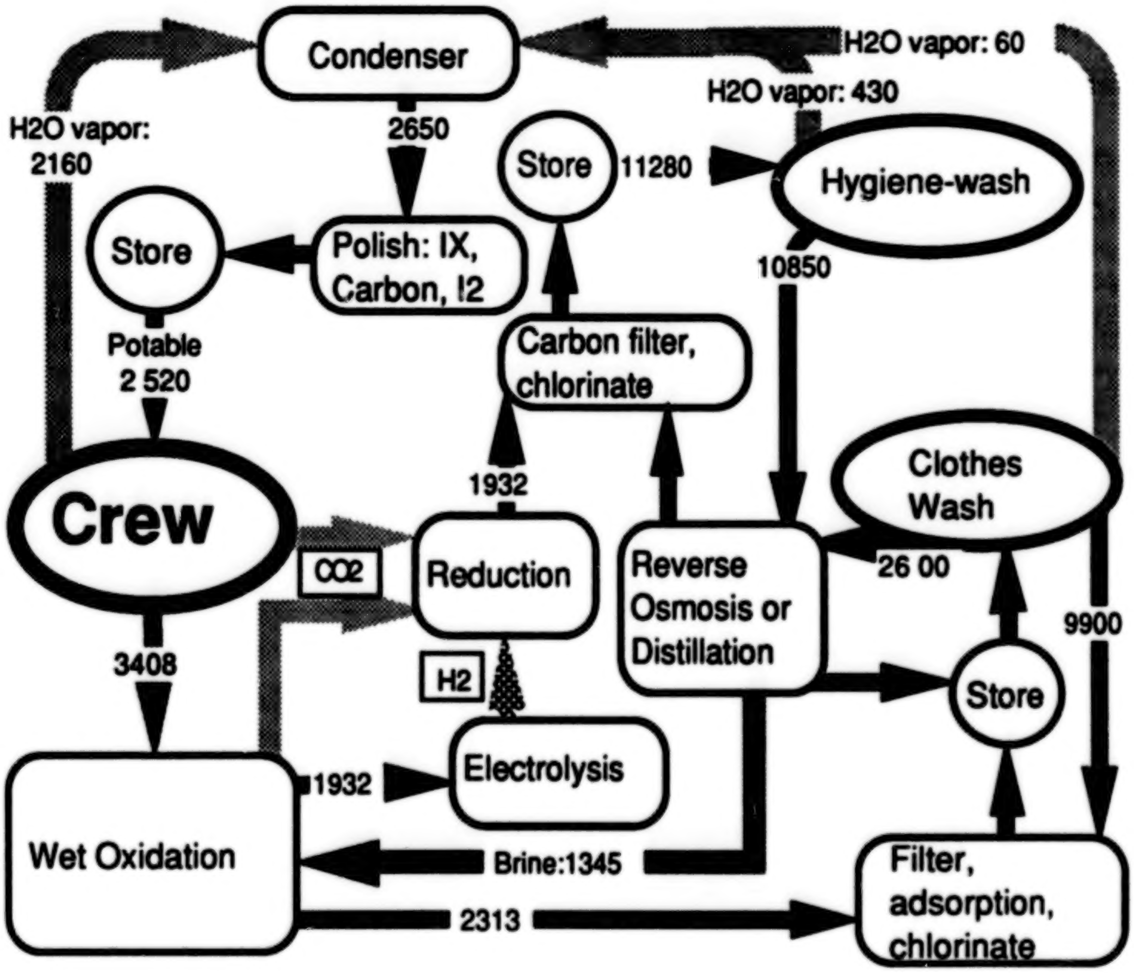

Condenser
H2O vapor: 60
H2O vapor: 430
H2O vapor: 2160
2650
Store
11280
Hygiene-wash
Store
Polish: IX, Carbon, I2
10850
Carbon filter, chlorinate
Potable 2 520
1932
Clothes Wash
Crew
CO2
Reduction
Reverse Osmosis or Distillation
26 00
9900
H2
3408
Store
1932
Electrolysis
Wet Oxidation
Brine:1345
Filter, adsorption, chlorinate
2313

Figure 1. Closed-loop water treatment example 1. Values in grams/person-day.

In this minimum example clothes-wash water is filtered, and organics may be partially removed through adsorption. Wet-oxidation water is of good quality, so it is added in the final rinse to compensate for evaporation losses. Electrochemical chlorination or ozonation is appropriate if this water is likely to be contaminated by bacteria. Monitoring conductivity is appropriate: a fraction of this water on exceeding 1000 ppM should be passed for purification to the RO (or distillation) process. This also makes up for evaporation losses in the hygiene-water cycle.

Spreadsheet analysis - This Example 1 and other performance examples can be tested for mass balance by spreadsheet analysis as in Table 2. The amounts of each source water (named in the first column) are estimated in column 2, passed through treatments in the next column, to product water. The percentage of total of estimated user needs is in the last column.

Source	**Grams/ person-day**	**Process** RO factors:	**Output1** 0.90	**Brine** 0.10	**Out TDS (initial)**	**(ppM) (final)***	**User Needs**
Clothes	2600	R O	2340	260	25	168	Potable
wash	9900	Filter	9900		116	730	2520
Hand+shower	5450	R O	4903	545	163	163	
wash							Semi-Potable
Dishes	5400	R O	4859	540	88	88	10896
wash							
Total Hygiene	23350		22001	1345	111	402	Wash Only
							12984
Urine	1568						
Urine flush	494	add Urine	2062		21996	22066	
Trash+Feces+	1346						
Urine Out+	4753	Wetox	4245	85	1309	1320	
RO Brines							
Cabin air	2650	condense	2650		160	160	
Humidities							Electrolysis
CO2 Redctn	818	Bosch	1932	includes	wetox.CO2		1932

Table 2. Wastewater sources, processing and water needs comparison. Values are grams per person-day. Water sources and use estimates from Wydeven and Golub (31) and references cited therein. RO and wet oxidation rejections are preliminary estimates.

(*) Concentration reached with recycle of 70% filtered, 30% RO, clothes-wash water, after ca. 100 cycles.

Simulations - Overall tests of the water reclamation system and estimates of constituent/contaminant flows and changes under various possible and extreme conditions may be made by computer simulations of the processing elements. The elements may then be linked to simulate an entire water cycle. Where products and starting compositions are known from tests, the simulation can be reviewed for accuracy of representation.

The humidity-condensation process was simulated first. The objectives were to estimate the contaminants expected in water condensed from air, the contaminants removed by this means, and the requirements for and best locations of activated carbon (in the inlet air stream, in the outlet air stream, or in the water stream).

The condensation process has as the input vapor: Air, temperature 22^{o} C, relative humidity 50% = 13 ppT water, and CO_2 at 400 ppM. A resulting water phase is the amount of condensate in equilibrium with the air: the latent water from cabin air, 2650 g/person-day (32). An estimate of the principal air contaminants expected in a space station (14) is given, for the principal constituents in descending order, in the first two columns of Table 3, for a crew of two (32). The calculated concentrations at a cabin flow of 25 m^3/hr are in the third column. A simulation calculation of the amount dissolving in the condensate water yields the contaminant concentrations of the next columns.

If the constituents are trapped by activated carbon from the air phase assuming a "ventilation" rate of 10 exchanges per day, the concentrations of the 6th column are obtained. At the rate identified by Otsuji and Yoshimura (5 exchanges per day), results in the 5th column are obtained. A more recent table is given by Leban and Wagner (8). This lists metabolic generation rates and Space Station Freedom generation rates, in comparison with Spacecraft Maximum Allowable Concentrations derived from occupational exposure values.

Constituent	Cabin-Air amount, per person of 2-person JEM (mg/day)	Air concentration, (ng/liter)	1/day	Water, after condensing (mg/liter) 5 exch.	10 exch.	MAC
Dichloromethane	1126.	3.75	218.	43.6	21.8	
Butanol-1	977.	3.26	412.	82.5	41.2	
Ethanol	654.	2.18	249.	49.9	24.9	
m-Xylene	629.	2.10	265.	53.0	26.5	10.
Methyl ethyl ketone	620.	2.06	215.	43.1	21.5	
Acetone	526.	1.75	139.	27.9	14.0	
Propyl acetate	336.	1.12	140.	26.7	13.4	
Methyl isobutyl ketone	258.	0.86	106.	21.2	10.6	
Propanol-2	254.	0.85	100.	20.0	10.0	
Ammonia	238.	0.79	86.	19.4**	9.9*	0.5
Carbon monoxide	219.	0.73	0.02	0.00	0.00	
Butanal	194.	0.65	66.	13.1	6.6	
Cyclohexane	194.	0.65	66.	13.3	6.7	
Toluene	170.	0.57	66.	13.6	6.8	2.
Cyclohexanol	164.	0.55	70.	14.0	7.0	
Total organics	6382.	21.3	2114.	421.9	211.	

*includes 9.1 as ammonium bicarbonate plus 0.61 as ammonium acetate. pH=8.1.

**includes 17.4 as ammonium bicarbonate plus 1.21 as ammonium acetate. pH=8.3.

Table 3. Simulated Contaminant Carry-over in Humidity-condensation Process. Data from Yoshimura (7) used in PROCESS simulation program.

Example 2.

Supercritical wet oxidation is a step that remains under evaluation. While a competing electrochemical oxidation process (33, 24) has also not yet been proved satisfactory for dependable use in space, for a second example of an oxidation step it makes an interesting other approach. It is of some additional interest in that hydrogen evolved from the non-oxidizing electrode can be used for the CO_2 reduction process, providing an economy of operations. Some ozone and/or chlorine may be evolved, of possible use in sterilizing final product water.

We will suppose that the process cannot be relied upon completely to provide potable or even semipotable water, but that sterility can be assured by the oxidizing electrode and that enough of the organic content of the washwaters and urine is destroyed so that product water can be returned as input to a reverse-osmosis unit. Some indestructible material will accumulate as brine recycling between the electrolysis and RO units, so a true waste stream is generated.

Removal of ionic waste is satisfied by an electrodialysis concentrating compartment included in the electrolysis cell. (The current density required for oxidation may exceed that allowed for electrodialysis, so this figure may be oversimplified.) The concentration of the ED brine is chosen by adjustment of the drain. This now multifunctional electrolysis cell is drawn in Figure 2.

Figure 3 is a diagram of this water-recycling configuration. Most of the elements are similar to those of the first example, except for restructuring the electrolysis module to include oxidizing and brine-concentrating functions. For this example the RO module has been increased in size to include all the clothes-wash wastewater, as another significant option.

SUMMARY AND RECOMMENDATIONS

A basic water recycling system to provide and assure quality of potable water, hygiene-wash water, and laundry water has been described. Assurance of performance relies on simplicity, redundancy, and in-flight repairability. Water-quality assurance relies on choices of source waters, redundant monitoring, and, through monitoring, a means for failure analysis.

The development route is clear, leading to qualification of all devices needed for assurance of water quality in a system for recycling 90% or more of spacecraft human needs. 100% of water requirements can be reached if water chemically bound in trash materials is recovered by a waste-oxidation process.

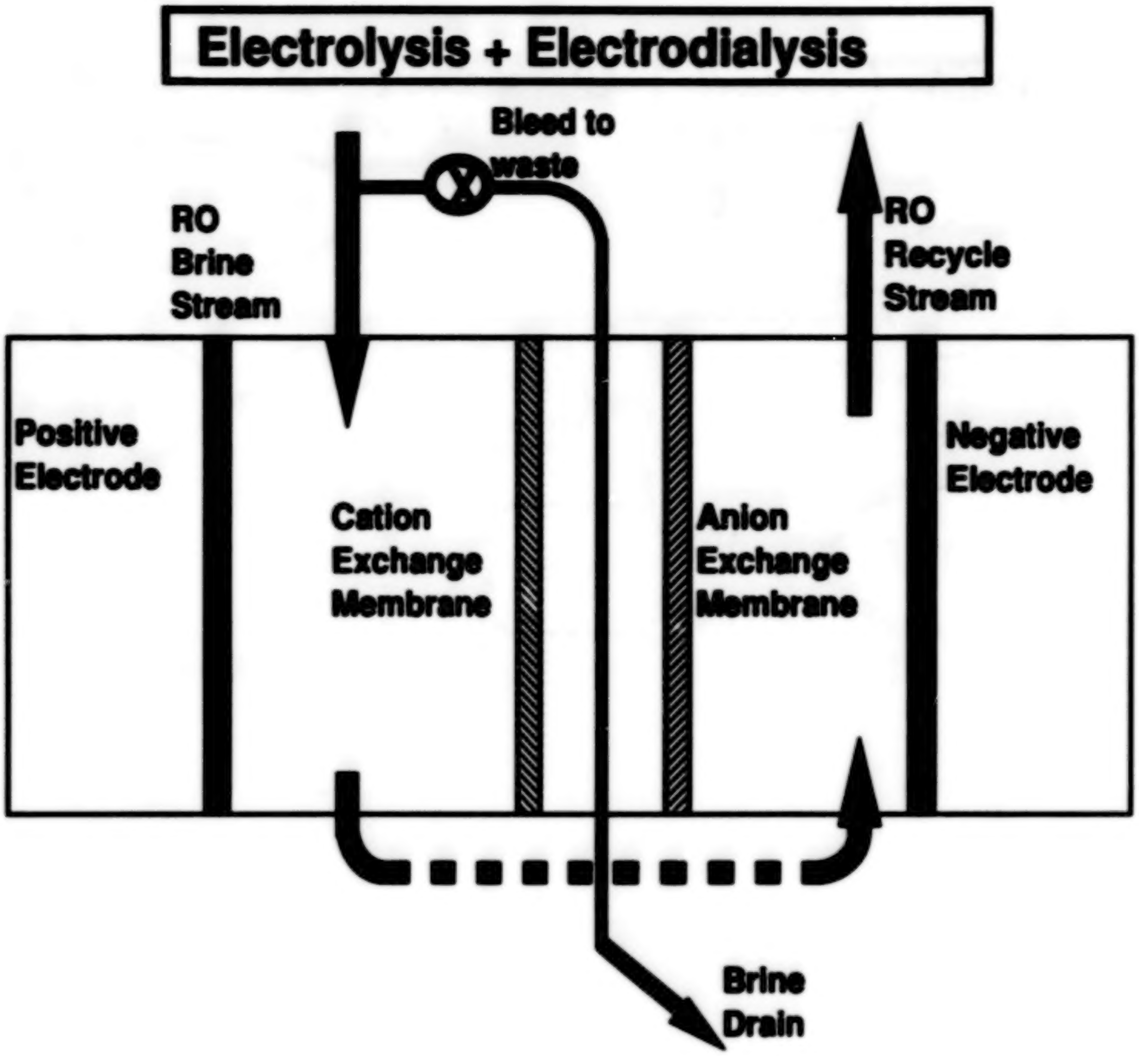


Figure 2. Combined Electrolytic Cell of Example 2. Organic contaminants of RO reject brine are supposed incompletely oxidized by the positive electrode: some gas (CO_2 and O_2) will be evolved and centrifugally separated. Some hydrogen will be evolved at the negative electrode, for use by the Bosch reduction process. Ionic constituents are further concentrated by passage through ion-exchange membranes, as conventional electrodialysis.

Critical decisions - Decisions necessary for specification of water recycling and water quality assurance equipment are in part externally determined by definitions of mission, crew requirements and construction materials, and in part by definitions of water-treatment processes. Spacecraft and habitat environmental temperature and pressure choices also influence the water cycle, and so are to be included in whole-system simulations. Several values external to the water-processing system, but influencing its performance, are required as input data to obtain realistic predictions of water quality as a function of process choices. These include: Accurate specification of water volume and time-of-use requirements for the mission, availability of waste heat, as from electronic components, medical specification of levels for additives to potable water: fluoride, calcium, bactericide.

Material choices of plastic or metal spacecraft components significantly influence the needs for contaminant-trapping. Final decisions on the size and type of contaminant traps require (1) that results of current toxicological research be followed for new information on hazards and limits, and (2) that components of the spacecraft that can directly or indirectly contact the water supply be characterized as soon as selected, for normal degassing and corrosion products and also decomposition and corrosion products of failure modes' (e.g. overheating) . It is not implied that materials should automatically be excluded for degradation potential, since in most cases adequate control of degradation products can be provided.

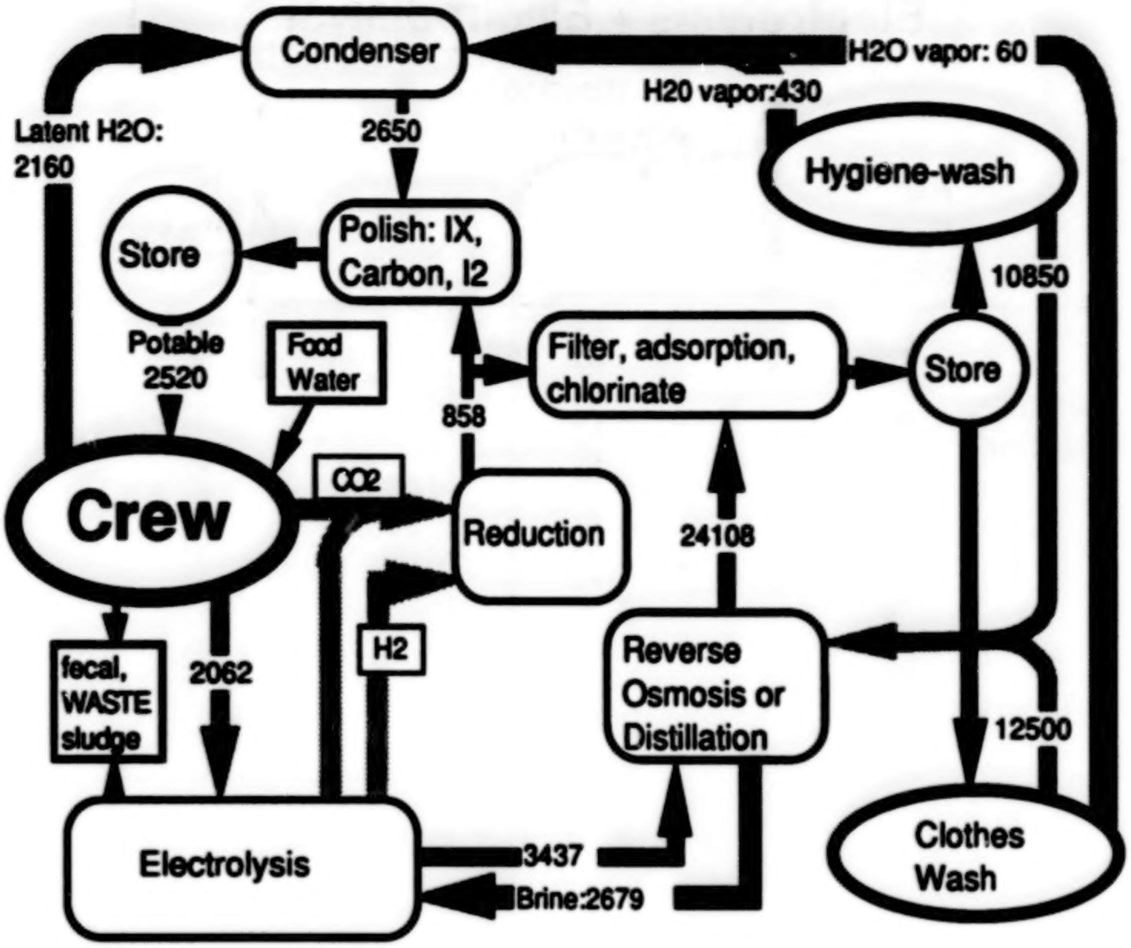

Condenser
H2O vapor: 60
H20 vapor:430
Latent H2O:
2160
2650
Hygiene-wash
Store
Polish: IX,
Carbon, I2
10850
Potable
2520
Food
Water
Filter, adsorption,
chlorinate
Store
858
CO2
Crew
Reduction
24108
H2
fecal,
WASTE
sludge
2062
Reverse
Osmosis or
Distillation
12500
Electrolysis
3437
Brine:2679
Clothes
Wash

Figure 3. Example 2: Closed-Loop Water Supply.

The most critical water-system decisions are seen as: oxidation method (supercritical wet oxidation vs. other), water purification (RO vs distillation), and impurity trapping (ion exchange and carbon adequacy). These choices are to be made on the basis of recommended testing.

The method applied here, analysis of wastewater sources and water needs, followed by choice and adaptation of technology, aand optimization using chemical process simulations by computer, can be applied to many industrial and home situations, where recycling can provide both safe disposal and reliable supply.

REFERENCES

1. Evanich, P. L.: Advanced Physical-Chemical Life Support Systems Research. Paper 881010, 18th Intersociety Conference on Environmental Systems, San Francisco 1988.
2. Noyez, G. P.: Carbon Dioxide Reduction Processes for Spacecraft ECLSS: a Comprehensive Review. Paper 881042, 18th Intersociety Conference on Environmental Systems, San Francisco 1988.
3. Lard, E. W.; Birnbaum, B. H.; and Deane, T. N.: Shipboard Laundry Wastewater Treatment Systems. Paper 76-ENAs-48, 6th Intersociety Conference on Environmental Systems, 1976.
4. Carrasquillo, R. L.: ARS/Resistojet Interfaces. In: ECLSS Working Group Meeting. Marshall Spaceflight Center, Aug. 24-26, 1988.
5. Gabbrielli, E. A.: Tailored Process for Remineralization and Potabilization of Desalinated Water. Desalination 39, 503 (1981).
6. Gabbrielli, E.; and Gerofi, J.P.: Appropriate Mineral Content of Desalination Water - Theory and Drinkers' Reaction. Desalination 49, 95 (1984).

7. Yoshimura, Y.; Manabe, K.; Kamishima, N.; Minemoto, M.; Hatano, S.; Etoh, T.; and Iida, H.: Study of Trace Contaminant Control System for Space Station. Paper 881117, 18th Intersociety Conference on Environmental Systems, San Francisco 1988.

8. Leban, M. I.; and Wagner, P. A.: Space Station Freedom Gaseous Trace Contaminant Load Model Development. Paper 891513, 19th Intersociety Conference on Environmental Systems, San Diego 1989.

9. Verschueren, K.: Handbook of Environmental Data on Organic Compounds. Van Nostrand Reinhold 1977.

10. Zdankiewicz, E. M.; and Price, D. F.: Phase Change Water Processing for Space Station. Paper 851346, 15th Intersociety Conference on Environmental Systems, San Francisco, July 1985.

11. Tleimat, B.; Laird, A. D. K.; and Howe, E.: Analysis and Cost Prediction of Reclaiming Agricultural Drainage Water Using Multieffect Vapor-Compression Distillation. Final Report to California Department of Water Resources, UC/DWR Agreement B-55037 Task Order 84-1, Nov. 1985.

12. Tleimat, B.; and Tleimat, M.: A Novel 2500 GPD 5-Effects Wiped-Film Rotating-Disk Vapor-Compression Module; Preliminary Results. Desalination 74, 289 (1989).

13. Verostko, C. E.; Garcia, R.; and Sauer, R.: Test Results on Reuse of Reclaimed Shower Water - a Summary. Paper 891443, 19th Intersociety Conference on Environmental Systems, San Diego 1989.

14. Jiang, Ji; Mingji, S.; Minling, F.; and Jiayan, C.: Study on the Interaction Between Membranes and Organic Solutes by the HPLC Method. Desalination 71, 107, 1989.

15. Haxo, H. E. Jr.; Miedema, J. A.; and Nelson, N. A.: Permeability of Polymeric Membrane Lining Materials for Waste Management Facilities. in: Migration of Gases, Liquids and Solids in Elastomers Symposium, American Chemical Society Rubber Division 126th Meeting, Denver, October 1984; Elastomerics 117, #5, 29 (1985).

16. Jordan, D.R.; Miyake, T.; and McIlhenny, W.F.: Brine Concentration by Electrodialysis, Phase 1. Phase 2: Jordan, D. R.; Bearden, M. D.; Komori, R.; and McIlhenny, W. F.: R & D Reports 74-930 and 74-931, Office of Saline Water, U.S. Department of the Interior, March 1974.

17. Klein, G.: Calculation of Ideal or Empirically Modified Mass-Action Equilibria in Heterovalent, Multicomponent Ion Exchange. Computers in Chemical Engineering 8, 171, (1984).

18. Herrmann, C. C.; Eaton, R. C.; Nyugen, C.; Martin, B.; and Klein, G.: Calcium Removal from Agricultural Drainage Water by Ion Exchange. Report to the State of California Department of Water Resources, Agreement B-55037, 1986.

19. Berger, B. B.: Control of Organic Substances in Water and Wastewater. Noyes Data Corp., 1987.

20. Kemmer, F. N.: The NALCO Water Handbook. Nalco Chemical Company, 1987.

21. Hong, G. T.; Fowler, P. K.; Killilea, W. R.; and Swallow, K. C.: Supercritical Water Oxidation: Treatment of Human Waste and System Configuration Tradeoff Study. Paper 871444, 17th Intersociety Conference on Environmental Systems, Seattle 1987.

22. Killilea, W. R.; Hong, G. T.; Shallow, K. C.; and Thomason, T. B.: Supercritical Water Oxidation: Microgravity Solids Separation. Paper 881038, 18th Intersociety Conference on Environmental Systems, San Francisco 1988.

23. Takahashi, Y.: Water Oxidation Waste Management System for a CELSS - the State of the Art. Biol. Sci. in Space 3, 45 (1989).

24. Bockris, John O'M.: Electrochemical Processing of Solid Waste. NASA CR-181128 (N87-25443), 1987.

25. Putnam, D. F.; Columbo, G. V.; and Michalek, W. F.: Pre- and Post-treatment techniques for spacecraft water recovery. Final Report, Contract NAS9-17073 (N87-25766).

26. Sauer, R. L.; Janik, D. S.; and Thorstenson, Y. R.: Medical Effects of Iodine Disinfection Products in Spacecraft Water. Paper 871490, 17th Intersociety Conference on Environmental Systems, Seattle 1987.

27. Bull, R. J.: Toxicological Aspects of Water Recycle and Disinfection. Paper 871491, 17th Intersociety Conference on Environmental Systems, Seattle 1987.

28. Gillies, M. T.: Potable Water from Wastewater. Noyes Data Corp. 1981.

29. Gnann, M.: Electrochemical Production of Highly Concentrated Ozone Using Lead Oxide Anodes. (PhD Thesis, Munich Technical University); NASA document N87-25530, 1985.

30. Smith, E.D.; and Scholze, R. J. Jr.: USA-CERL's Wastewater Recycle R&D Program. Paper 871520, 17th Intersociety Conference on Environmental Systems, Seattle 1987.

31. Wydeven, T.; and Golub, M.: NASA Technical Memorandum in press (1990).

32. Otsuji, K.; Hanabusa, O.; Sawada, T.; Satoh, S.; and Minemoto, M.: An Experimental Study of the Bosch and the Sabatier CO_2 Reduction Processes. Paper 871517, 17th Intersociety Conference on Environmental Systems, Seattle 1987.

33. Putnam, D. F.: Composition and Concentrative Properties of Human Urine. Report DAC-61125-F1 on contract NAS1-8954, McDonnell Douglas Aircraft Company for NASA Langley Research Center, June 1970.

WATER QUALITY ANALYZER

Warren Kelliher
Aerospace Technologist
NASA Langley Research Center
Hampton, VA 23665

DOCUMENTATION OF THIS PAPER WAS NOT PROVIDED FOR INCLUSION IN THESE PROCEEDINGS. FOR FURTHER INFORMATION, PLEASE DIRECT ALL INQUIRIES TO THE NAME AND ADDRESS LISTED ABOVE.

N91

24053

UNCLAS

N91-24053

NEW RESEARCH ON BIOREGENERATIVE AIR/WATER PURIFICATION SYSTEMS

Anne H. Johnson
Science and Technology Laboratory
National Aeronautics and Space Administration
John C. Stennis Space Center
Stennis Space Center, MS. 39529-6000

R.D. Ellender
Paul J. Watkins
Department of Biological Sciences
University of Southern Mississippi
Box 5153
Hattiesburg, MS. 39606

INTRODUCTION

For the past several years, the Science and Technology Laboratory at Stennis Space Center (SSC) has been involved in the development and application of air and water purification systems. This technology is based on the combined activities of plants and microorganisms as they function in a natural environment. Early efforts dealt with the use of artificial or constructed wetlands for wastewater treatment. Numerous communities as well as corporations have adopted this technology. In fact, all of the wastewater at SSC is treated using these types of systems. More recently, researchers have begun to address the problems associated with indoor air pollution. Various common houseplants are currently being evaluated for their abilities to reduce concentrations of volatile organic compounds (VOCS) such as formaldehyde and benzene.

With development of the Space Exploration Initiative (SEI), there will be significant increases in mission duration. Problems with resupply necessitate implementation of regenerative technology. Although the final system may primarily be based on physicochemical processes, it is feasible to consider the application of bioregenerative technology for the air/water purification.

Aspects of bioregenerative technology developed at SSC have been included in a prototype habitat known as the BioHome (Figure 1). A 650 SF structure, the BioHome functions as a pilot system to facilitate analysis of bioregenerative technology in a semi-closed environment. The ultimate goal is to employ this technology in conjunction with physicochemical systems for air and water purification within closed systems.

The BioHome is divided into two regions. one is designated as a living area while the second contains the wastewater treatment system. This system is a modified version of an artificial wetland, relying on vascular plants and microorganisms to effect the treatment process. The system is housed within 6 - 8 inch segments of polyvinylchloride (PVC) pipe ranging in length from 10 to 12.75 ft. and contains plants such as canna lilies (Canna) and bulrush (Scirpus). In addition, there are various types of porous substrate included such as activated carbon. Due to increased surface area, the substrate material promotes biofilm development, a process integral to successful treatment of wastewater (1). In addition, biofilms also play a role in the presence or absence of bacterial pathogens (2).

Prior to inclusion of bioregenerative air or water purification systems in a closed environment, it is necessary to fully assess the associated risks. It is expected that wastewater will have a characteristic microflora, some of which will be pathogenic. Similarly, biological contaminants may be airborne. The bulk of the latter group will probably originate in the abundant plant material present. There is a potential problem in closed systems with build-up of airborne microbes that may be attributed to the lack of ozone and ultraviolet rays. These elements are present outdoors and comprise what is known as the "open air factor" (3). Consequently, there is a tendency for microbial survival to be enhanced indoors due to the absence of this effect. Devices such as HEPA filters may be used to reduce some biological airborne contaminants, however they will not alleviate the problem. Similarly, chemical contaminants may occur in ambient air. They stem from a variety of sources including building material, plants, and humans.

Earlier preliminary studies have dealt with partial assessments of biological contaminants in the BioHome. Data indicated that the wastewater treatment system exhibited tremendous potential for reduction of bacterial pathogens such as Salmonella (97.53%) and Shigella (98.52%) (4). Similarly, the biological oxygen demand (BOD) and fecal coliform counts were significantly reduced (Tables 1, 2). Studies analyzing ambient microflora revealed relatively low levels of bacteria and fungi present. Bacterial genera included Bacillus, Escherichia, Flavobacterium, Klebsiella, Micrococcus, and Staphylococcus. Fungal isolates were identified as members of the genera Aspergillus, Mucor, and Penicillium.

The purpose of this study was to continue the risk assessment of bioregenerative technology with emphasis on biological hazards. In an effort to evaluate the risk for human infection, analyses were directed at enumeration of fecal streptococci and enteric viruses within the BioHome wastewater treatment system.

MATERIALS AND METHODS

Fecal Streptococci Analysis: For a period of ten weeks, weekly water samples were taken from both effluent (segment 1) and effluent (segment 6) sites of the treatment system. Using the membrane filtration technique, appropriate volumes of sample were analyzed using Gelman GN-6 0.45 m sterile filters. Following filtration, the filters were aseptically transferred to KF agar and incubated at 35°C for 48 hours (5). The density of fecal streptococci/enterococci per 100 ml was calculated using only those plates with colonies numbering in the desired range (20 to 60). Verification of isolates was accomplished according to the protocols outline in A.P.H.A.'s Standard Methods (5).

Enteric Virus Analysis: Measured quantities of wastewater were pumped through 90 mm 1MDS Virosorb membranes for a total of 27 samples. The majority of samples were taken from the effluent sampling port. Additional samples were obtained from segments 3 and 4 as well as the septic tank. 90 mm membranes were eluted using 80 ml of 0.1 M glycine, pH 10.5. The eluent from this step was then passed through a 47 mm Virosorb membrane and eluted with 5 ml of 0.1 glycine, pH 10.5. Next, 10% PSF and 0.1 (10X) gentamicin was added, sample pH was adjusted to 7.0, then the sample was incubated at 35°C for one hour. Samples were then centrifuged at 1900 X g for 20 minutes, filtered (0.20 micron), and distributed into 1.5 ml aliquots for storage at -70°C. For purposes of inoculation, Linbro plates were prepared from stock MA-104 cells and allowed to settle for 24 hours. Next, the growth medium (L-15) was removed by aspiration and each monolayer inoculated with 0.1 ml of undiluted sample. Following an incubation period of one hour at 35°C, monolayers were covered with 1 ml of maintenance medium and incubation continued. Plates were observed daily for evidence of cytopathic effect for a total of seven days (6).

RESULTS

Results of the fecal streptococci analysis indicated that the wastewater treatment system significantly reduced numbers of this group (Table 3). Influent samples over the 10 week period averaged 53 CFUs (colony forming units)/100 ml. None of the effluent samples exhibited any growth. Consequently, the system is 100% effective in the reduction of fecal streptococci/enterococci.

To date, no viruses have been isolated from any portion of the wastewater system. 27 samples were screened for the presence of enterics with no evidence of cytopathic effect.

DISCUSSION

It is encouraging to find that fecal streptococci are virtually removed from the wastewater. This group, also known as the Group D streptococci, has been linked to high incidences of urinary tract infections as well as abdominal lesions and are resistant to numerous antibiotics (7). Similarly, the absence of enteric viruses is a promising finding. There are several factors that may account for the low levels and/or absence of these groups. First, the septicity of the tank preceding the artificial wetland may be such that conditions are unfavorable for both groups. Factors such as high NH_3 content may limit survival, particularly with respect to enteric viruses. It is also possible that the relative numbers of both groups are comparatively low in the raw wastewater. The majority of sewage that is used for the BioHome studies is derived from that which is generated on site at SSC. Consequently, the presence or absence of a particular group of microorganisms is a reflection of resident microbial population associated with the raw wastewater.

The analysis of these data along with previous studies support the finding that artificial wetlands may provide a suitable means of reducing the number of pathogens in wastewater (8, 9). Several studies have documented the advantages of aquatic and wetland plants for the treatment of wastewater (10, 11). It has been theorized that plants perform two functions in an artificial wetland system. The first is that they provide increased surface area for microbial attachment, an important consideration since the treatment process relies on microbial activity. The second function relates to the transport of oxygen to the root zone, or rhizosphere, thereby producing an aerobic environment (12). The resultant aerobic zone supports a microbial consortium that effects modification of nutrients, ions, and other compounds while the aerobic/anaerobic interface serves to enhance the processes of nitrification and denitrification (13).

It is interesting to note that plants have additional mechanisms to dictate the types of microorganisms found within the rhizosphere. Studies by Bowen and Rovira (14) revealed that several regions of the root produce compounds that leak from the root or may be pumped out as a result of metabolic activity. Such compounds were identified as inhibitory to certain microorganisms. Broadbent et al (15) theorized that such antibiotic activity may be involved in significant coliform reductions associated with artificial wetlands. Similarly, Palmateer et al (16) found that coliform reduction was enhanced substantially during the summer. This reduction coincided with an anoxic period, suggesting the ability of plants to translocate oxygen to the rhizosphere, thereby providing an explanation for improved coliform removal in vegetated systems.

These findings are also supported by Seidel (17) whose studies included Juncus effusus, Scirpus lacustris and Phragmites communis. Seidel maintains that excretions from the plants either partially or completely kill pathogenic bacteria while heterotrophs are left unharmed. Unfortunately, the author neglected to include relevant reference material. Consequently, the validity of these finds must be carefully considered.

Pathogens are known to be removed by physical/chemical processes (filtration and adsorption) and by biological inactivation and predation (18). However, biofilm development also plays an important role in their presence or absence. In a study utilizing granular activated carbon (GAC) , it was determined that the autochthonous microbial community influenced pathogen survival (2). When pathogens were introduced to sterile GAC in the presence of heterotrophs, they attached at levels similar to those found in pure culture, then decreased. However, when the two were added to GAC with a mature biofilm, the pathogens attached at lower levels and decreased at a more rapid rate.

Future research will address the enumeration of bacterial pathogens as it relates to biofilm development on activated carbon. similarly, efforts will continue in the characterization of fecal streptococci and enteric viruses associated with the wastewater treatment system.

ACKNOWLEDGMENT

This work was supported by funding from the Technology Utilization program, National Aeronautics and Space Administration.

LITERATURE CITED

1. Antonie, R. L.: Fixed Biological Surfaces - Wastewater Treatment. CRC Press, Inc., (West Palm Beach), 1978.

2. Camper, A. K.; LeChevallier, M.W.; Broadaway, S. C.; and McFeters, G.A.: Growth and Persistence of Pathogens on Granular Activated Carbon Filters. Appl. Environ. Microbiol. vol 50, 1985, pp. 1378-1382.

3. Cox, C. S.: The Aerobiological Pathway for Microorganisms. John Wiley and Sons (New York), 1987.

4. Johnson, A. H.; Bounds, B. K.; and Gardner, W.: I. Assessment of Internal Contamination Problems Associated with Bioregenerative Air/Water Purification Systems. SAE Technical Paper Series, Proceedings of the 20th Intersociety Conference on Environmental Systems, Williamsburg, VA., July 9-12, 1990.

5. A.P.H.A. Standard Methods for the Examination of Water and Wastewater, 17th Edition, 1989.

6. Block, J. C.; and Schwartzbrod, L.: Detection and Identification of Viruses in Water Systems. VCH Publishers (New York), 1989.

7. Brock, T. D.; Smith, D. W.; and Madigan, M. T. Biology of Microorganisms. 4th Edition. Prentice-Hall, Inc., (Englewood Cliffs, NJ), 1984.

8. Gersberg, R. M.; Gearheart, R. A.; and Ives, M.: Pathogen Removal in Constructed Wetlands. In D. A. Hammer (ed.), Constructed Wetlands for Wastewater Treatment: Municipal, Industrial and Agricultural, pp. 431-445. Lewis Publishers, Inc., (Chelsea, MI), 1989.

9. Gersberg, R. M.; Lyon, S. R.; Brenner, R.; and Elkins, B. V.: Fate of Viruses in Artificial Wetlands. Appl. Environ. Microbiol. vol 53, 1987, pp. 731-736.

10. Hammer, D. A.; and Bastian, R. K.: Wetland Ecosystems: Natural Water Purifiers? In D. A. Hammer (ed.), Constructed Wetlands for Wastewater Treatment: Municipal, Industrial and Agricultural, pp. 2-19. Lewis Publishers, Inc., (Chelsea, MI), 1989.

11. Kadlec, J. A.: Nutrient Dynamics in Wetlands. In K. R. Reddy and W. H. Smith (ed.), Aquatic Plants for Water Treatment and Resource Recovery, pp. 393-419. Magnolia Publishing, Inc., (Orlando), 1987.

12. Armstrong, W.: Oxygen Diffusion From the Roots of British Bog Plants. Nature vol 204, 1964, pp. 801-802.

13. Good, B. J.; and Patrick, W. H.: Root-Water-Sediment Interface Processes. In K. R. Reddy and W. H. Smith (ed.), Aquatic Plants for Water Treatment and Resource Recovery, pp. 359-372. Magnolia Publishing, Inc., (Orlando), 1987.

14. Bowen, G. D. ; and Rovira, A. D. : Microbial Colonization of Plant Roots. Annu. Rev. Phytopath. vol 12, 1976, pp. 181-197.

15. Broadbent, P.; Baker, K. F.; and Waterworth, Y.: Bacteria and Actinomycetes Antagonistic to Fungal Root Pathogen in Australian Soils. Aust. J. Biol. Sci. vol 24, 1971, pp. 925-944.

16. Palmateer, G. A.; Kutas, W. L.; Walsh, M. J., and Koellner, J. E.: Abstracts of the 85th Annual Meeting of the Am. Soc. for Microbiol. Las Vegas, NV, March 3-7, 1985.

17. Seidel, K.: Macrophytes and Water Purification. In J. Tourbier and R. W. Pierson, Jr. (ed.), Biological Control of Water Pollution, pp. 109-121. University of Pennsylvania Press (Philadelphia), 1976.

18. Gersberg, R. M.; Brenner, S. R.; Lyon, S. R.; and Elkins, B. V.: Survival of Bacteria and Viruses in Municipal Wastewater Applied to Artificial Wetlands. In K. R. Reddy and W. H. Smith (ed.), Aquatic Plants for Water Treatment and Resource Recovery, pp. 237-246. Magnolia Publishing, Inc., (Orlando), 1987.

TABLE 1
BIOHOME MEAN MONTHLY BOD VALUES

DATE	INFLUENT (mg/L)	EFFLUENT (mg/L)
6/89	368	1.2
7/89	264	2.0
8/89	217	20.6
9/89	388	3.8
10/89	293	7.2
11/89	304	2.0
12/89	245	2.2
1/90	114	1.8
2/90	234	7.2
3/90	236	1.5
4/90	224	11.6
5/90	357	3.1

TABLE 2
BIOHOME MEAN MONTHLY FECAL COLIFORM COUNTS

DATE	INFLUENT	EFFLUENT
6/89	4.8×10^5	1
7/89	1.5×10^5	1
8/89	8.0×10^6	800
9/89	4.4×10^6	6000
10/89	8.5×10^5	8000
11/89	4.0×10^4	1
12/89	8.0×10^5	6000
1/90	4.2×10^4	530
2/90	8.0×10^4	10
3/90	8.0×10^4	1
4/90	2.5×10^4	150
5/90	8.0×10^5	1

TABLE 3
BIOHOME FECAL STREPTOCOCCI DENSITIES (CFUs/100 ml)

WEEK #	INFLUENT	EFFLUENT
1	58	0
2	57	0
3	59	0
4	45	0
5	56	0
6	50	0
7	57	0
8	49	0
9	50	0
10	52	0

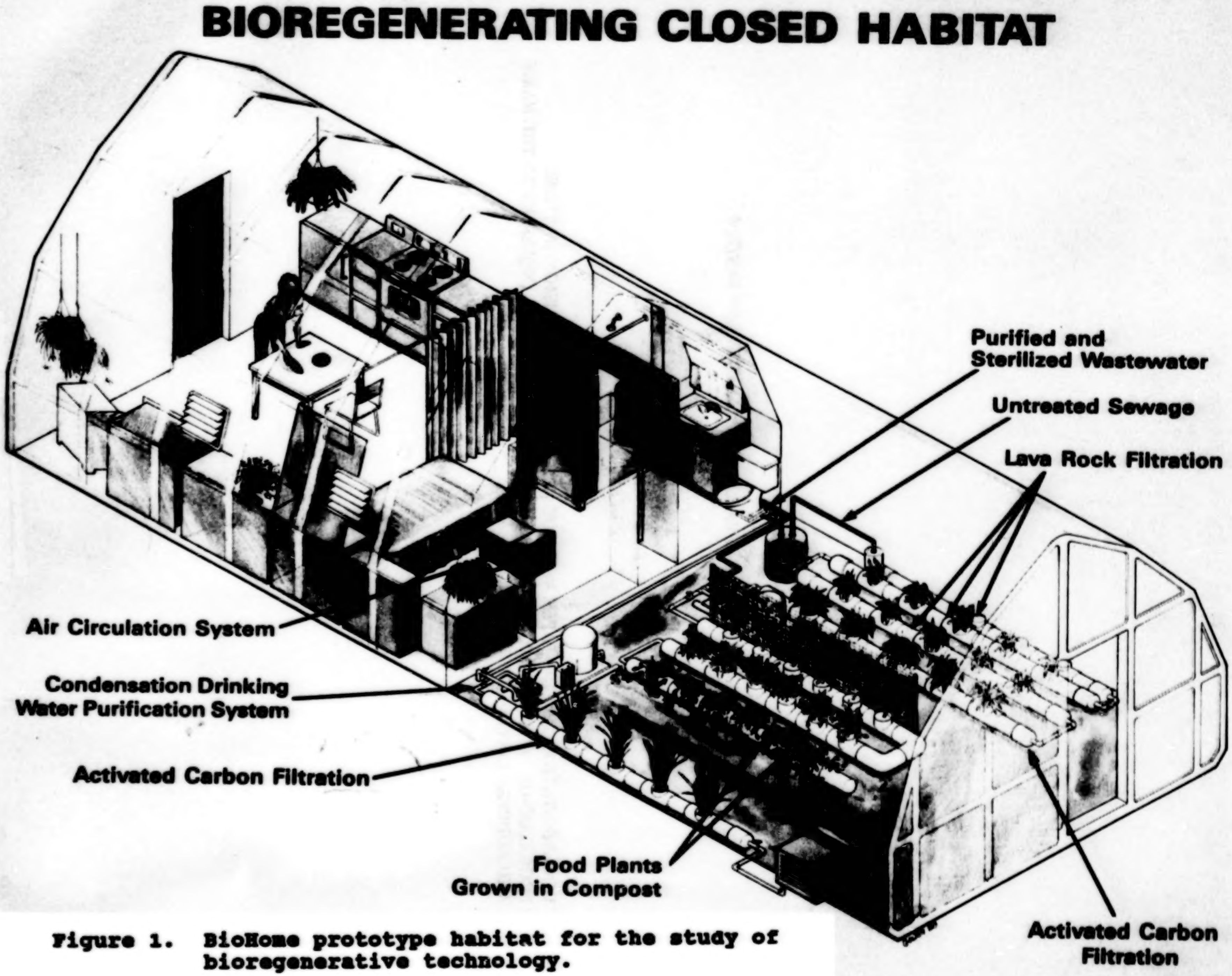

BIOREGENERATING CLOSED HABITAT
Purified and Sterilized Wastewater
Untreated Sewage
Lava Rock Filtration
Air Circulation System
Condensation Drinking Water Purification System
Activated Carbon Filtration
Food Plants Grown in Compost
Activated Carbon Filtration

Figure 1. BioHome prototype habitat for the study of bioregenerative technology.

ENVIRONMENTAL AND FACILITIES MANAGEMENT SYSTEM

Bruce Davis
Geographer
Stennis Space Center
SSC, MS 39529

DOCUMENTATION OF THIS PAPER WAS NOT PROVIDED FOR INCLUSION IN THESE PROCEEDINGS. FOR FURTHER INFORMATION, PLEASE DIRECT ALL INQUIRIES TO THE NAME AND ADDRESS LISTED ABOVE.

N91

24054

UNCLAS

N91-24054

THE LAND ANALYSIS SYSTEM (LAS)

Yun-Chi Lu
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

Fred M. Irani
Stephen W. Wharton
ST Systems Corp.
Lanham, Maryland 20706

ABSTRACT

The Land Analysis System (LAS) is an interactive software system, available in the public domain, for the analysis, display, and management of multispectral and other digital image data. The system was developed to support the earth sciences research and development activities at the NASA Goddard Space Flight Center and the USGS EROS Data Center. Initially released in the fall of 1985, LAS provides over 240 applications functions and utilities, a flexible user interface, complete on-line and hardcopy documentation, extensive image data-file management, reformatting, and conversion utilities, and high-level device independent access to image display hardware. This paper summarizes the capabilities of the latest release of the system (Version 5.0). Emphasis is given to the system portability and the isolation of hardware and software dependencies in this release.

INTRODUCTION

The Land Analysis System (LAS) was developed by the Space Data and Computing Division and the Laboratory for Terrestrial Physics at the NASA Goddard Space Flight Center (GSFC) in cooperation with the U.S. Geological Survey EROS Data Center (EDC). The Transportable Applications Environment (TAE) [1,2] is used to provide a standard user interface and a unified environment for function development[1]. The LAS was originally implemented under the Digital Equipment Corporation[2] VMS operating system and installed on several VAX 11/780 computer systems at GSFC and EDC following the successful completion of acceptance testing [3] in August 1985. Since that time, the number of LAS installations has increased to over 80 sites to support research, production and education oriented image processing. A description of the initial release of LAS (version 3.2) can be found in an earlier paper [4].

The immediate objective in developing LAS was to satisfy the image processing requirements as identified by the science user community. The long term objectives were to: 1) develop transportable software that can be shared among government agencies and installed on a variety of computer systems; 2) incorporate a comprehensive user-friendly interface and array of executive services supporting raster data input, management, and display; and 3) provide a reliable, fully documented, and easily supported library of applications software functions that can be combined to support routine operational tasks, while maintaining a flexible environment in support of research for earth science data applications.

---

[1] LAS and TAE are available in the public domain through the Computer Software Management and Information Center (COSMIC) at the University of Georgia, Athens, Georgia.

[2] Any use of trade names and trademarks in this publication is for descriptive purposes only and does not constitute endorsement by NASA or the U. S. Geological Survey.

There is an ongoing effort to further enhance the functionality and utility of LAS, not only through the development of new and enhanced applications functions and system utilities but also through the implementation of LAS on microprocessor based workstations running under non-VMS operating systems. The extension of LAS to the workstation domain provides a number of advantages: 1) allows greater flexibility in the configuration of hardware and software to meet the needs of a particular application; 2) lessens the potential impact of processing resource conflicts on users by placing a significant amount of dedicated processing resource directly under the control of the user; and 3) allows more efficient utilization of the host's computer's resources by distributing highly interactive data capture, display, and analysis tasks to the workstation. Also, the availability of LAS is expanded to a larger number of potential users.

APPROACH

To maximize the portability of LAS it was necessary to minimize and isolate dependencies on specific computer hardware, operating system, and peripheral devices. These portability requirements have had a major influence on the continuing evolution of LAS by dictating the development of highly transportable and functionally modular software as well as the utilization of hardware independent interfaces and integration techniques. The utilization of TAE inherently enhances the portability of LAS, isolating the user from the underlying operating system and assisting the applications programmer by providing such services as a common data interface for terminal access and communication, parameter definition and ingest, message handling, and session logging. In addition, TAE provides considerable flexibility in packaging software to meet the needs of a particular application by providing convenient mechanisms for user menu definition and allowing commonly used functions to be combined into single procedures or "procs." The availability of these services frees program developers from redundant design tasks and simplifies the development and integration of new applications functions.

The three major enhancements to the system which were requested by NASA users [5] will be incorporated in the Version 5.0 release. Details of these enhancements are presented in this paper. Two of the three enhancements are already complete: 1) removal of the file group implementation (in which each band is stored in a separate file) of the multispectral images; and 2) developing a portable implementation of the system. The third enhancement, improved catalog manager performance in file archival and restore, will be completed for the 5.0 release in December 1990.

SYSTEM OVERVIEW

The LAS provides a flexible and comprehensive environment for image processing by integrating the following software system components:

- the TAE user interface that is consistent, flexible, and can be used effectively by both novice and experienced users from an ASCII terminal or by mouse-driven commands on a graphics work-station;
- a comprehensive set of over 240 applications functions for input, analysis, and storage of a variety of digital image and geographic data in raster form;
- a uniform syntax for file naming spanning different operating systems;
- the Display Management Subsystem (DMS) to provide high level device independent manipulation of color image displays and 8-bit work-stations under X-Windows;
- interface routines to support data transfer between the LAS and optional, site-specific external software packages (e.g., geographic data entry and statistical analysis systems); and
- support routine libraries and documentation to facilitate the development of new applications functions.

The LAS is fully documented with hardcopy manuals and on-line help information under TAE. Image history and session logs are automatically maintained at the discretion of the user. Processing history information is maintained for all LAS images and includes image names, applications functions used, and parameter values. The user can elect to add comments to, list or delete records in the image history. The optional session log is kept for interactive sessions by TAE. It is a record of the pertinent user entries, program messages, and results appearing at the user's terminal. The user also has the option to print or delete the current session log before logging off the system. The tabular data interface provides flexibility in that appropriate interface functions may be modified to use different data base management, statistical analysis, and vector data processing systems. In addition, system manager documentation and release notes are provided for installing, running, and maintaining the system.

OVERALL DESIGN ENHANCEMENTS

The current release of LAS contains significant improvements to the previous release (Version 4. 1). The design and implementation of the entire system has been substantially changed while maintaining as much as possible the external appearance, functionality and operability of the original system. Many of the design changes were made in the interest of enhanced hardware and operating system portability. Moreover, the subroutine libraries were redesigned to provide simplified scenarios for applications programmers. The myriad special file formats of Version 4.1 were eliminated in favor of more system uniform file formats. In addition, an effort was made to use the ASCII file format wherever appropriate to allow users to directly create, list, edit or transmit such files over network and other distributed system environments.

A set of LAS coding standards for the C and FORTRAN languages was developed for the development of transportable software. The LAS library of system support routines for disk and tape access and data manipulation, originally written in the VMS MacroAssembler language was rewritten in C. The majority of the system is now implemented in the C programming language although there are still few components of the system written in the standard FORTRAN-77 language. The Image Input/Output subroutines were entirely rewritten to be based on the TAE "xi" image I/O package. A side-effect of this decision is an increased dependence on the port of TAE to a new operating system before LAS can be implemented in that same environment. All bands of an image are now contained in a single file (rather than a separate file for each band as in earlier LAS versions). This configuration allows applications to either process all image bands at once as if the data were stored in a band interleaved-by-line format or one band at a time, as deemed most appropriate and efficient by the applications programmer. Processing all bands at once has significantly improved wall-clock performance benchmarks for most LAS applications.

Image files (denoted by the suffix ".img") have logically associated history (suffix ".his") and data descriptor record (suffix ".ddr") files which are used during image manipulation to determine image size, data-type and other specialized information related to the image. Image geo-referencing is now carried throughout the LAS system by use of new entries in the ".ddr" files. Applications such as TM data magnetic tape ingest, geometric transformation, image concatenation and others now access the new fields in the ".ddr" files created for that purpose. For example, ingest of a TIPS format full scene from magnetic tape now places the latitude and longitude coordinates for the four corners of the image into the disk image's data descriptor record at the time of data ingest. The "lab.lab" file, of the LAS Version 4.1 and earlier, used the VMS exclusive Indexed Sequential Access Method (ISAM) format. This VMS dependency has been eliminated by the image associated files ".img", ".ddr" and ".his".

The Baseline library (i.e., the LAS support subroutines and the applications functions) has been built and tested under VAX/VMS on several different Digital Equipment Corporation computers, and under UNIX-based operating systems on SUN3 and SUN4 work-stations and a Powernode 9050 computer. Specific non-portable software modules are identified for potential users so that estimates of the level of effort required to port to new machines can be made more readily. Areas in which adjustments are known to be required for new system ports are in the area of the magnetic tape support library where operating system interfaces are necessary to manipulate magnetic tape drives and in many C language include files which define constant values such as the maximum and minimum range of values for data types such as float, long, etc.

Such areas as may need modification for untested operating systems are limited to the support software libraries rather than the application functions themselves with the exception of one or two specialized application functions which must behave differently depending on the host operating system.

The Baseline library can be used under any of the above named operating systems. The Contributed library is a collection of unsupported software from the LAS user sites that is available as part of the system release. The baseline version of the release has been configured into four separate directory trees called TAE, World, LAS and DMS. The LAS and DMS directories contain the unique applications and support modules which make up each system. The World directory consists of Applications and support routines which are required by both LAS and DMS users and programmers. This structure was created to avoid maintaining duplicate software for LAS or DMS users who may be interested in installing only one of these systems on their host computer. The TAE directory which is included in the release consists of the latest version of TAE currently installed at the user site and may be upgraded independently of the LAS, DMS, and World directories.

FUNCTIONAL ENHANCEMENTS

User Interface

User access to the LAS software is provided by the Transportable Applications Environment (TAE). The TAE was developed as part of a separate project at GSFC as a generic user interface and Environment for applications programs. The TAE provides: menu or command-driven function selection; an option for interactive prompting of function parameters; on-line help capabilities; interactive asynchronous and batch processing; and access to the host operating system without exiting from TAE. Additionally, users now have the option of using the system via keyboard on traditional ASCII terminals (using TAE Classic) or by mouse-driven commands on graphic workstations (using TAE Facelift).

Under TAE-Classic or Facelift, the LAS functions can be selected from menu screens or by command entry. Function parameters can be specified interactively using the tutor screens or by command entry. The functions can be selected from menus or by command entry and can be executed interactively, asynchronously or in batch mode. In the menu mode a user can search through a hierarchical structured menu tree to locate a specific function. The menu choices are either applications functions, commands, help, or other menus. In the command mode, the user can activate a function by entering its name and the required parameters. Under TAE-Facelift users can select application programs or open help windows for applications via mouse-button clicks. The user may switch from menu mode to command mode at any time during a session and TAE-Facelift users can enable or disable Facelift at any time from command mode. For a more complete description of the TAE see [1,2].

When an application is selected in menu mode, TAE automatically displays the tutor screens to prompt the user for the needed parameters. The user may also enter the tutor screens by typing "TUTOR" and the function name while in command mode. The TAE-Classic user enters values by typing the parameter name and its value e.g., "IMAGE=Harrisburg" and may obtain additional information on any parameter by typing HELP and the parameter name e.g., "HELP IMAGE". Most parameters may be abbreviated to the set of characters that makes them unique, e.g., if no other parameters had "I" as the first letter then "IMAGE" could be abbreviated as "I". Once all required values are given, the user may either execute the function or save the parameter values to a parameter file. The parameter file can be recalled, edited, and used to execute single LAS functions in batch mode. Alteratively, a procedure file can be created using the host editor to execute a series of functions in batch mode.

The TAE-Classic provides a number of additional capabilities that serve to minimize the amount of repetitive typing needed to execute a LAS function: parameters are validated prior to program execution and the user is re-prompted for invalid entries; spectral and spatial subsets can be given as part of the image name; meaningful default values are provided for optional parameters; function sub-commands limit the prompts in tutor mode to only those parameters that are relevant under the selected processing option.

On-line help information is available to describe TAE menus, TAE commands, LAS applications programs, and program parameters.

TAE-Facelift users are able to reduce typing commands to a minimum while retaining the conveniences and functionality of the TAE-Classic mode of operation. By means of mouse-button selections, graphics-workstation users can select parameter options and open several help and utility windows simultaneously while tutoring on the selected application. Input file names and other such entries must still be typed at the keyboard, of course, but the amount of keying is greatly reduced by the ability to select most parameter values from option lists and save or restore parameter selections, run the application and perform other TAE operations by mouse-button clicks alone.

Applications Functions

The LAS includes over 240 applications functions to process raster data in a number of data types, including byte, integer*2, integer*4, and real*4. The documentation for each function includes: revision date, purpose, parameters, algorithm, error messages, examples, prerequisites, processing limits (e.g., maximum image size, number of categories, and data types allowed), and references. The hard-copy documentation is generated from the on-line documentation and is accessible through the TAE help facilities. The original functionality of the LAS Version 4.1 was preserved or enhanced in response to user requests during the implementation of the Version 5.0 technical standards.

File Management

A simplified file management capability has replaced the catalog manager sub-system of the earlier LAS design. This, along with other system enhancements, has reduced the load on system management, run-time resources and wall-clock response time overall. The new system still provides the basic functions of the original system without the overhead in system resources and need for special attention by the system manager. The system now provides: 1) naming conventions that are independent of the host operating system; 2) an alias capability for the abbreviation of file names; and 3) utilities for the archival and retrieval of data files to and from magnetic tape.

Display Management Subsystem

The LAS display application functions use the Display Management Subsystem (DMS) to provide a device-independent interface to color image display devices. DMS minimizes application software dependencies on a particular display device by providing a generic set of image manipulation functions for device allocation, deallocation, and initialization; image transfer and setup; image viewing alteration; cursor manipulation; and image enhancement. At the lowest level, DMS is dependent on the characteristics of the particular display being used. The DMS libraries above this level can be considered to be transportable, meaning they are not dependent on the capabilities of a particular display. This transportability also applies to the display application functions that use DMS. Consequently, when a new display is to be used, only the lowest level, device dependent libraries must be modified. Like TAE, DMS can be used independently of LAS. For a more detailed description of DMS see Perkins et al., [6].

LAS image data can be displayed and manipulated using special hardware features such as panning and zooming. The modified image, image intensity mapping tables or graphics, and annotation information can be saved in disk files for later use. The LAS display functions use the DMS utilities to: allocate and deallocate a display; show display status information; load images to the display; save display images to disk; assign colors to intensity values; apply, modify, save, and restore graylevel and pseudocolor mappings; flicker series of images and look-up tables; perform operations on images including arithmetic, logic, rotations, zoom, and convolutions; define, enable, or disable the cursor, and return intensity values and coordinates at cursor position. The mensuration and graphics display functions allow the placement and editing of points, lines, polygons, and annotation on the graphics planes of the display. These graphics can be associated with a specific image (or image subset), such as training sites for supervised classification.

DMS has undergone extensive modifications with the result that the new subroutine libraries are no-longer compatible with earlier versions. An X-Window version of device dependent software has been developed which allows limited display capabilities on 8-bit graphic work-stations. This capability may be used for quick-look, pseudo-coloring, and control-point selection. The X-Windows device-dependent routines were developed under the MIT X11R3 protocol and are therefore compatible with release 3 and 4 of the X11 software.

The GSFC version of LAS will contain the DMS Version 3.0 as it's display system while the DMS Version 1.2 and earlier will be available from the contributed software library as a part of the release. Display applications will be documented as LAS Display Modules (LDM) in the LAS Version 5.0. The "Umbrella" version of the display applications will be released, however, such that all display modules will be run under a controlling, menu-driven, application module. This scenario provides speed in operation but at the cost of flexibility in inter-mixing DMS and LAS commands. The nonumbrella version of the applications will also be made available to users through the contributed software library. The Display Management Subsystem under TAE provides a high level, device-independent interface to display hardware. DMS uses the low-level libraries for display manipulation as provided by the respective vendors including vendor implementation of the X-Windows graphics protocol. Thus, DMS allows new display hardware to be used with LAS without modification to the applications functions.

Support Routine Libraries

A set of support routine libraries were developed for use by applications programmers. The libraries include: geometric coordinate translation, image input/output, data descriptor record (ddr) manipulation, pixel value manipulation, statistics and magnetic tape I/O, and others. The geometric coordinate translation routines are used to translate coordinate pairs between different map projections. The image I/O functions provide efficient read, write, and update access to image data independent of data type. The data descriptor record manipulation (ddr) functions are used to read, write, or update the LAS image labels. Pixel manipulation functions perform unary (e.g., logarithm), binary (e.g., arithmetic or logical), and miscellaneous (e.g., data type conversion) operations on image lines or line segments. The statistics I/O functions can be used to update, search, and delete information within the hierarchy of the LAS statistics files. A utility library provides miscellaneous capabilities such as for parsing input image names, printing messages to users and others.

CONCLUSIONS

The LAS is used at EDC, GSFC, and over 80 other sites to support production, research, and education oriented image processing and Version 5.0 satisfies the user requirements specified in [5]. The current system design allows the same body of LAS source code to be built (e.g., compiled and linked) under both VMS and Unix-based operating systems. The new release has been tested under the VMS, SUN3, SUN4, Gould Powernode 9050 and IBM AIX operating systems. Dependencies on proprietary software have been removed while system interfaces to statistical and other outside software packages have been preserved. Other enhancements include a simplified file management system, enhanced display capabilities, and operations through mouse or keyboard.

The functional advantages of the current release (version 5.0) can be summarized as follows:

- The system is no longer bound to the VMS operating system. The same body of source code can be installed under VMS or UNIX-based operating systems.
- The Catalog Manager has been removed from the system greatly reducing system overhead and increasing overall response time and replaced by the File Management System which provides a uniform file naming syntax independent of the host operating system.
- Multi-band images have been consolidated into a single file, eliminating the file group

construct of previous versions of LAS.

- Hard copy applications users guides are generated from the online documentation. Online and offline documentation is now identical.
- Single user interface, under TAE independent of the host operating system in which novice users can select functions from menus and be prompted for parameters, while experienced users can select functions and define parameters via terse commands on ASCII terminals or by mouse driven commands on graphic work stations.
- The LAS, DMS, WORLD, and TAE support libraries facilitate the integration of new applications functions.
- The Display Management Subsystem under TAE provides high-level device independent access to image displays and 8-bit workstations under X-Windows.
- Tabular data interface routines in the LAS Contributed Library facilitate data transfer between LAS and site-specific external software systems

ACKNOWLEDGMENTS

A large number of organizations have been active in the LAS development with too many individual participants to be listed here. Therefore, the authors wish to collectively acknowledge the contributions to this project by the following organizations: NASA/GSFC Space Data and Computing Division, NASA/GSFC Laboratory for Terrestrial Physics, U.S. Geological Survey EROS Data Center, Science Applications Research, ST Systems Corp., Century Computing Corp., and Computer Sciences Corp.

LITERATURE CITED

1. Szczur, M.R., D.C. Perkins, and D.R. Howell, "T.A.E.: Transportable Applications Executive, NASA's Front-end for Scientific/Engineering Programs," Proc. of the Digital Equipment Computer Users Society, New Orleans, Louisiana, 1985, pp. 433-439.

2. Overview (TAE Transportable Applications Environment) Plus Version 4. 1, January 1990: TAE User Support Office, NASA/GSFC, Greenbelt, MD, 20771.

3. Kovalick, W.M., J.A. Newcomer, and S.W. Wharton, "A Methodology for Evaluation of an Interactive Multispectral Image Processing System," Photogrammetric Eng. and Remote Sensing, Vol. 53, No. 8, 1987, pp. 1087-1092.

4. Wharton, S.W., Y.C. Lu, B.K. Quirk, L.R. Oleson, and J.A. Newcomer, "The Land Analysis System (LAS) for Multispectral Image Processing," IEEE Trans. on Geoscience and Remote Sensing, 26(5), 1988, pp. 693-697.

5. Wharton, S.W., and Y.C. Lu, "The Land Analysis System (LAS): A General Purpose System for Multispectral Image Processing," IEEE International Geoscience and Remote Sensing Symposium, Ann Arbor, Michigan, 1987, Vol. H, pp. 1081-1086.

6. Perkins, D.C., M.R. Szczur, J. Owings, and R.K. Jamros, "A Device-Independent Interface for Image Display Software," 1984 Machine Proc. of Research Institute, Redlands, California, 1985, 430 pages.

N91

24055

UNCLAS

N91-24055

POTENTIAL COMMERCIAL USES OF EOS REMOTE SENSING PRODUCTS

Leslie L. Thompson
NASA/Goddard Space Flight Center
Code 925
Greenbelt, Maryland 20771

ABSTRACT

The instrument complement of the EOS satellite system will generate data sets with potential interest to a variety of users who are, for instance, now just beginning to develop geographic information systems tailored to their special applications and/or jurisdictions. Other users may be looking for a unique product that enhances competitive position. The generally distributed products from EOS will require additional value added processing to derive the unique products desired by specific users. Entrepreneurs have an opportunity to create these proprietary level 4 products from the EOS data sets. Specific instruments or collections of instruments could provide information for crop futures trading, mineral exploration, television and printed medium news products, regional and local government land management and planning, digital map directories, products for third world users, ocean fishing fleet probability of harvest forecasts, and other areas not even imagined at this time. This paper looks at the projected level 3 products that will be available at launch from EOS instruments and speculates about commercial uses of the data after value added processing.

BACKGROUND

The Earth Observing System is planned to the primary NASA system that supports the national and international Global Change Program. The first satellite platform will be launched in calendar year 1998. This platform will carry a payload of between 8 to 12 sensor systems. The selection of the instruments has not been made at this time. It may be by the time of the meeting at which this paper is being presented. The payload will be determined based on contributions to global science observations and financial constraints. The financial constraint depends on congressional action on the FY91 budget and total program authorization. Since the FY91 budget is still in negotiation between the President and Congress the payload cannot be selected at this time.

Over the past year an Investigators Working Group (IWG) has been meeting. The IWG has been active in providing recommendations to NASA management on the EOS Science Plan, establishing a preliminary list of data products to be available at launch, and recommending an instrument payload for the first EOS platform. It is the list of initial data products that this paper wants to make the end user community aware. At this time the number of at launch products approaches 3000 in number. This list will probably undergo evolution to be reduced to a set of products numbering in the 100's in order to be more manageable.

The data products of course depend on the instrument payload of the first EOS platform. The next section provides a short description for the candidate instruments. These instruments are considered to be Facility Instruments that support a large science team or Principle Investigator Instruments that support science in specific research areas. If you do not want to read the instrument descriptions skip over the next two sections.

The Proposed USA Facility Instruments for EOS-A

The following list contains those instruments that are currently considered the Facility instruments for EOS-A:

1) The Moderate Resolution Imaging Spectrometers-Nadir or Tilt (MODIS-N, MODIS-T)

The Nadir MODIS provides 24 hour day and night global coverage of the Earth by scanning through the nadir with a +- 55 degree field-of-regard. This instrument provides a nominal 1000m GIFOV from a 705km altitude in 38 narrow spectral bands from 413nm to 14.3um. Selected spectral bands will have improved ground sample distance to provide 8 bands with 500m GIFOV and 2 bands with 250m GIFOV. MODIS-N has removed the polarization sensing bands from its complement of capabilities. This instrument delivers 1.7 x 10^{12} bits per day to the Data and Information System.

The Tilt MODIS provides a visible spectral region spectrometer with a slightly narrower field-of-regard, +/- 45 degrees, and has a 1.1km GIFOV. The minimum spectral interval provides 15nm bandwidths over 32 contiguous bands starting at 410nm wavelength. The field-of-regard is tilted to positions fore and aft of the orbital track to avoid specular glint off of the oceans and to perform bidirectional reflectance studies. The ocean science for this instrument requires that polarization sensitivity be less than 2.5% over the field-of-regard for pointing angles less than 20 degrees fore or aft. The primary mission for MODIS-T is to support world bioproductivity (ocean color) research, thus bidirectional reflectance studies occur on a scheduled basis rather than all the time.

2) Atmospheric Infrared Sounder (AIRS)

AIRS is a scanning radiometer with a +/- 49 degree field-of-regard and provides continuous soundings of the atmosphere in up to 4280 contiguous spectral intervals (wavenumber interval between 0.5 cm-1 to 2.0 cm-1 in bandwidth) in the 3.7um-17um spectral region. AIRS will provide important atmospheric correction data for the MODIS in addition to its sounding products. The GIFOV equals 50km in most of the channels and selected channels with 15km GIFOV'S. This instrument has the potential to provide pressure and temperature profiles of the atmosphere with a vertical resolution of 1km. The current AIRS approach uses a grating spectrometer and requires a cryogenic refrigerator for the large infrared focal plane.

3) High Resolution Imaging Spectrometer (HIRIS)

HIRIS compliments MODIS operation by providing higher spatial resolution with a narrower swath. At nadir the GIFOV is 30m with a swath width of 30km. This instrument operates as a system to sample selected data about the spacecraft subtrack. The FOV can be directed to targets within a field-of-regard 60 degrees up track, 30 degrees down track, and +/-24 degrees acrosstrack. The HIRIS is an imaging grating spectrometer with spectral ranges in the visible to near IR and the SWIR. As with AIRS the infrared focal plane is cryogenically cooled with a refrigerator.

4) Advanced Microwave Sounding Unit (AMSU)

AMSU is a microwave radiometer providing measurements of atmospheric temperature and humidity. This complements the infrared measurements of AIRS. It is a 20-channel instrument divided into AMSU-A and AMSU-B subsystems. AMSU-A primarily provides atmospheric temperature from the surface to 40km altitude in 15 channels, i.e., 23.8 GHz, 31.4 GHz, 12 channels between 50.3 to 57.3 GHz, and 89 GHz. Coverage is approximately 50 degrees on both sides of nadir, with a GIFOV of 50km. AMSU-B provides atmospheric water vapor in 5 channels at 89 GHz, 166 GHz, and 183 GHz. Coverage is the same as AMSU-A, but the GIFOV is 15km.

Currently, AMSU-A is the only AMSU subsystem planned for EOS-A, although the accommodation of AMSU-B is being studied.

Potential Principle Investigator Instruments for EOS-A

The following Principal Investigator instruments may be included in the EOS-A payload:

1) Cloud and the Earth's Radiant Energy System (CERES)

The CERES are a pair of scanning radiometers similar to the Earth Radiation Budget instruments flown on previous missions. They are broadband radiometers with three channels: total radiance (.0.2um to 100um), shortwave (0.2um to 3.5um), and longwave (6um to 25um). One radiometer scans crosstrack, and the second scans in a constant angle rotating plane. The GIFOV is 2.3 degrees by 3.4 degrees.

2) High Resolution Microwave Spectrometer Sounder (HIMSS)

HIMSS is a dual band scanning microwave radiometer. It scans in a constant angle rotating plane with a GIFOV of 5km at 90 GHz and a 50km GIFOV at 6.6 GHz. HIMSS is used to derive parameters for precipitation rate, cloud water, water vapor, sea surface roughness, ice, and snow. HIMSS has a 2m parabolic antenna with a scanning feed subsystem.

The following PI instruments are new capabilities with characteristics that compliment the facility instruments:

3) Multi-angle Imaging Spectroradiometer (MISR)

MISR provides data on bidirectional reflectance properties using fixed field-of-view pushbroom sensors operating in the visible to near infrared. There are four sensors looking forward and four looking aft with a separate nadir viewing pushbroom sensors. The viewing angles are 28.5°, 45.6°, 60°, and 72.5°. Four narrow spectral bands are used in each of the sensors in the spectral region between 440nm and 860nm. The spatial resolution is 1.73km in normal mode and 216m in local mode. The field-of-view is 27° for the sensors observing at an angle of 28.5°. This swath width is matched by the other sensors with the FOV corrected for the angle off the nadir. All FOV's are centered on the orbital track.

4) GPS Geoscience Instrument (GGI)

The GGI uses 18 dual frequency Global Positioning System (GPS) receiver-processor units and 3 distributed GPS antennas to provide centimeter level global geodesy, atmospheric temperature profiling, ionospheric gravity wave detection and tomographic mapping, and precise attitude determination in support of other instruments.

5) Lightning Imaging Sensor (LIS)

The Lightning Imaging Sensor observes the distribution and variation of lightning over the Earth. This sensor uses a solid-state staring army with a very narrow band filter in the visible spectrum and special signal processing electronics to threshold and detect lightning events and the time of their occurrence.

6) Earth Observing Scanning Polarimeter (EOSP)

The EOSP simultaneously measures radiance and degree of polarization in 12 spectral bands from 410nm to 2250nm. The GIFOV is 10km over a swath of +/- 55 degrees acrosstrack. This is the only polarimeter proposed for EOS.

7) High Resolution Dynamic Limb Sounder (HIRDLS)

HIRDLS is a synergistic combination of two PI instruments to provide a single capability to observe atmospheric constituents in the upper troposphere, mesosphere and stratosphere. This radiometer scans in azimuth and elevation while observing the atmosphere above the limb of the earth. 3km vertical steps are observed over a 70km range in altitude above the limb. 12 spectral bands are used from 6.1um to 17.3um.

8) Measurements of Pollution I in the Troposhere (MOPIT or Tropospheric Radiometer for

Atmospheric Chemistry and Environmental Research (TRACER)

These instruments measure the upwelling radiation from carbon monoxide and potentially methane and ammonia in the troposhere. Both use gas correlation radiometers and differential detection processing. One or the other of these sensors are candidates for EOS-A.

9) Stick Scatterometer (STIK SCAT)

This is an active microwave radar that uses 6 fixed fan beam antennas and operates in the 13.995 GHz regime with dual polarization. This instrument uses a doppler processor and provides 1200km swaths with 25/50 km resolution.

Potential Japanese Facility Instrument for EOS-A

The Intermediate Thermal Infrared Radiometer (ITIR) is a candidate to be supplied by the Japanese Space Agency (NASDA). The ITIR is a 14 band pushbroom scan radiometer with 60km a crosstrack swath. There are 3 visible bands at 15m GIFOV, 6 SWIR bands at 30m GIFOV, and 5 thermal infrared bands at 90m GIFOV. The field-of-view is pointable both fore and aft and in the crosstrack.

PROPOSED DATA PRODUCTS

This paper will not try to give an detailed preview of EOS data products that are planned to be available at launch and then in the time frame a year post launch. However, the nature of the data base is shown.

The proposed list of data products for the EOS program now numbers around 3300 for both the at launch and post launch time frames. The at launch products will be comprised mostly of Level 1 radiometrically corrected digital data and Level 2 radiometric data that has been gridded to one of several Earth located map projections. Of more interest will be Level 3 products that begin to derive geophysical parameters. Many Level 3 products will be available at launch. These Level 3 products will encompass ocean chlorophyll, sea surface temperature, land surface temperature, fire locations, cloud cover, vegetation index, vegetation type, biomass, atmospheric temperature and humidity profiles, net energy balance, total ozone, precipitation rate, snow cover and water content, and many others. It is the Level 3 products which can provide an information source that receives value added processing that leads to products for end users. EOS participating scientists will be generating algorithms for a large range of Level 4 products, but an expanding end user community may need the Level 4 product tailored to their specific situation or may need a unique product developed from Level 2 and 3 data. It is in the area of value added processing that commercial opportunities exist to develop and distribute proprietary products to different public and private users.

In order to provide a flavor the data product data base that is being created for EOS by Dr. Al Fleig of NASA Goddard Space Flight Center, Tables 1 and 2 are shown as examples of what is currently collected into the books. At some time in the future after decisions are made to reduce the overlap and redundancy between various investigators, a more compact list of products will be available for consideration.

Table 1 illustrates the data as collected by common data product area. The first column just list the sequential number of the product in the data base. The second column (CAT) identifies category of earth science research; in this example the "AR" is for Atmospheric Radiation. The third column (Parameter) is self explanatory. The fourth column (Type) indicates that the investigator is Interdisciplinary (II), Facility (FI), or Principal (PI). The fifth column (Investigator) is the name of the scientist involved. The sixth column (Instr.) is the related EOS instrument. The seventh column (Units) provides the relevant measurement parameters. The eighth column (Accuracy) is self explanatory, but perhaps a little cryptic with respect to percentages of what. The next three columns provide the anticipated temporal, horizontal, and vertical resolutions for the data product (note: N/A is not applicable). The last column (Time Frame) shows that the product will be available At Launch, Post Launch, or not known at this time.

Prod #	Cat	Parameter :: Qualifier	Type	Investigator	Instr.	Units	Accuracy Abs :: Rel	Temporal Resolution	Horizontal Resol. :: Domain	Vertical Resol. :: Domain	Time frame
2051	AR	Cloud_cover ::	I :: II	Barron		%	5 :: 5	1 dy	30 m :: Local	N/A :: Cloud	
2052	AR	Cloud_cover ::	I :: II	Hansen			3% ::	1-wk avg	500 km :: Global	:: Cloud	
2053	AR	Cloud_cover ::	I :: II	Isacks		%		1/wk avg	5 km :: Land/R	N/A :: Cloud	
2054	AR	Cloud_cover ::	I :: II	Lau		%	5% :: 5%	12 hr	50 km :: Regional	N/A :: Atmos	
2055	AR	Cloud_cover ::	I :: II	Liu					:: Ocean	N/A :: Cloud	
2056	AR	Cloud_cover ::	I :: II	McNutt			5% ::		:: Canada/R	N/A :: Cloud	
2057	AR	Cloud_cover ::	I :: II	Moore		% cover	10% :: 10%	wk	1 km :: Global		
2058	AR	Cloud_cover ::	I :: II	Murakami		% cover	10% ::			N/A :: Cloud	
2059	AR	Cloud_cover ::	I :: II	Sellers				4/dy	100 km ::	0.5 km :: Trop	
2060	AR	Cloud_cover ::	I :: II	Srokosz		%	5% :: 1%	2/dy	10 km :: Ocean/S.A	N/A :: Cloud	
2061	AR	Cloud_cover ::	I :: II	Wielicki			2% :: 2%	6/dy(d,n)	25-100 km :: Global	N/A :: Atmos	
2062	AR	Cloud_cover ::	O :: FI	Chahine,Smith	AIRS/AMSU	dimensionless	0.05 :: 0.025	2/dy(d,n)	15 x 45 km :: Global	N/A :: Cloud	AL
2063	AR	Cloud_cover ::	O :: FI	Smith	AIRS/AMSU	dimensionless	0.05 :: 0.025	2/dy (d,n)	15 x 45 km :: Global	N/A :: Cloud	AL
2064	AR	Cloud_cover ::	O :: II	Barron		%		1 dy	10 km :: Regional		
2065	AR	Cloud_cover ::	O :: II	Barron		%		5 min	30 km :: East.U.S		
2066	AR	Cloud_cover ::	O :: II	Barron		%		5 min	2 km :: East.U.S		
2067	AR	Cloud_cover ::	O :: II	Wielicki		fraction	5% :: 1%	18/dy(d,n)	25 km :: Regional	:: Atmos	
2068	AR	Cloud_cover :: Areal_extent	O :: FI	Kaufman	MODIS-N	km^2		mo	1 dg :: Global	N/A :: Sfc	FL
2069	AR	Cloud_cover :: Cirrus	I :: II	Bates		/m/sr		1 dy-avg	100 km :: Global	0.5 km :: Trop	
2070	AR	Cloud_cover :: Cirrus	I :: II	Lau		%	5% :: 5%	1 dy	100 km :: Global	N/A ::	
2071	AR	Cloud_cover :: Cirrus	O :: FI	Hardesty	LAWS	/m/sr		1 dy-avg	100 km :: Global	0.5 km :: Trop	AL
2072	AR	Cloud_cover :: Fraction	I :: II	Bates		dimensionless	0.05 :: 0.025	2/dy(d,n)	15 x 45 km :: Global	N/A :: Cloud	
2073	AR	Cloud_cover :: Fraction	I :: II	Bates			:: 10%	6 hr	1 x 1 dg :: Global	N/A :: Cloud	
2074	AR	Cloud_cover :: Fraction	I :: II	Bates		%	10% :: 5%	dy,mo	1 dg :: Global	N/A :: Cloud	
2075	AR	Cloud_cover :: Fraction	I :: II	Kerr,Sorooshian		%	5% :: 5%	1 dy	10 km :: Land/R	N/A :: Cloud	
2076	AR	Cloud_cover :: Fraction	I :: II	Rothrock		dimensionless	0.1 :: 0.1	1 dy	100 km :: Polar	N/A :: Cloud	
2077	AR	Cloud_cover :: Fraction	I :: II	Wielicki			2% :: 2%	16 dy	30 m :: Regional	N/A :: Atmos	
2078	AR	Cloud_Cover :: Fraction	O :: FI	Spinhirne	GLRS	%	1% ::	2-16 dy	10-200 km :: Global	N/A ::	AL
2079	AR	Cloud_cover :: Fraction	O :: FI	Welch	HIRIS	dimensionless	1% :: 0.5%	1-3 min,2-16 dy	30 m :: Local	:: Cloud	AL
2080	AR	Cloud_cover :: Fraction	O :: FI	Welch	ITIR	dimensionless	3% :: 3%	16 dy	15-30 m :: Local	N/A :: Cloud	AL
2081	AR	Cloud_cover :: Fraction	O :: FI	King	MODIS-N	%	10% :: 5%	2/dy(d,n), mo	5 km :: Global	N/A :: Cloud	AL
2082	AR	Cloud_cover :: Fraction	O :: FI	King	MODIS-N	%	10% :: 5%	1 dy,mo	1 x 1 dg :: Global	N/A :: Cloud	AL
2083	AR	Cloud_cover :: Fraction	O :: II	Bates		dimensionless		20 min	50 km :: Global	N/A :: High_cloud	
2084	AR	Cloud_cover :: Fraction	O :: II	Bates		dimensionless		20 min	50 km :: Global	N/A :: Mid_clouds	
2085	AR	Cloud_cover :: Fraction	O :: II	Bates		dimensionless		20 min	50 km :: Global	N/A :: Low_clouds	
2086	AR	Cloud_cover :: Fraction	O :: FI	Barkstrom	CERES	dimensionless	5% :: 2%	6/dy (d,n)	25 km :: Global	N/A :: Atmos	AL
2087	AR	Cloud_cover :: Fraction	O :: FI	Barkstrom	CERES	dimensionless	5% :: 2%	6 hr	1.25 x 1.25 dg :: Global	N/A :: Atmos	AL
2088	AR	Cloud_cover :: Fraction	O :: FI	Barkstrom	CERES	dimensionless	5% :: 2%	dy-,mo-avg	1.25 x 1.25 dg :: Global	N/A :: Atmos	AL
2089	AR	Cloud_cover :: Fractional	O :: II	Barron		fraction		12 hr	4.5 x 7.5 dg :: Global		
2090	AR	Cloud_cover :: Fractional	O :: II	Barron		fraction		12 hr	2.8 x 2.8 dg :: Global		
2091	AR	Cloud_cover :: OH	I :: II	Hansen			3% ::	1-wk avg	500 km :: Global	:: Trop	
2092	AR	Cloud_cover :: Perimeter	O :: FI	Kaufman	MODIS-N	km		mo	1 dg :: Global	N/A :: Sfc	FL
2093	AR	Cloud_cover :: Size_distribution	O :: FI	Welch	ITIR	dimensionless		16 dy	15-90 m :: Local	N/A :: Cloud	AL
2094	AR	Cloud_JPDF ::	O :: FI	King,Menzel	MODIS-N&T	dimensionless		1 dy , mo	1 dg :: Global	N/A :: N/A	FL
2095	AR	Color_index :: Soil	O :: FI	Huete	MODIS-N&T	class	10% :: 5%	mo	1 km :: Land/R	N/A :: Sfc	AL
2096	AR	Cross_section :: Altimeter	I :: II	Srokosz		dB	0.2dB :: 0.1dB	10 dy	10 km :: Ocean/S.A	N/A :: Sfc	
2097	AR	Cross_section :: Backscatter	I :: II	Brower		dB	10% :: TBD	dy,seas	25 km :: Ocean	N/A :: Sfc	
2098	AR	Cross_section :: Backscatter	I :: II	Cihlar			2 dB :: 1 dB	3 mo	25 m :: Canada/R	N/A :: Sfc	
2099	AR	Cross_section :: Backscatter	I :: II	Cihlar			2 dB :: 1 dB	3 mo	25 m :: Canada/R	N/A :: Sfc	
2100	AR	Cross_section :: Backscatter	I :: II	Cihlar			2 dB :: 1 dB	6 mo	10 m :: Canada/R	N/A :: Sfc	

TABLE 1: EXAMPLE OF EOS DATA PRODUCTS IN THE SCIENCE OF CLOUDS

It is evident from Table 1 that significant overlap does exist between investigators and their proposed products. It is probable that each instrument team will have one person responsible for a single product. Products common between instruments will provide cross correlation and verification of each others product. The Interdisciplinary Investigators may come to depend on the instrument learns for many of these overlap products. Thus, there is the expectation that the number of products can be seduced as indicated earlier.

Table 2 is an illustration of products anticipated for the MODIS instrument. MODIS has a very broad range of utility and will be a primary source of data for many products related to global change.

Having provided these examples of the data base, it is now time to speculate regarding potential commercial uses.

SPECULATIONS ON THE COMMERCIAL USES OF EOS DATA

Several years ago the book Megatrends projected that we would increasing depend on information technology. Judging by the large number of articles on geographic information systems now appearing in the journal Photogrammetric Engineering & Remote Sensing, both hardware and software have progressed to be able to support multidimensional data overlays. It is in this environment that it seems likely that EOS Level 3&4 data will find potential commercial application. Government data policy will impact this potential to some extent. With 8 years to launch, data access policy and pricing can be worked between the government and the science and end user communities.

The business community that will supply the value added processing to tailor EOS data products to the needs of their customers will supply much of the insight and imagination required to provide the commercial products. ne next set of paragraphs suggest some areas that have already been in consideration. Cost to the user and lack of timeliness have limited the success of these endeavors in the current time frame.

Commodity Futures Trading

Global information on crop type, estimates of acreage undercultivation, and crop vigor may be a data set that can be derived from EOS data. Data on vegetation index, crop cover type, areal extent of cultivation, precipitation, soil water content, and long term climate trends are expected to be products provided on a global scale by EOS passive electro-optical sensors (MODIS, AIRS, & CERES), and passive microwave sensors (AMSU & HIMSS). Some clever individuals may find a way to combine the essential elements in a unique approach that is beneficial to futures trading in commodities or to farming interests that want to optimally compete in the global market. The Goddard Space Flight Center work by Dr. Jim Tucker in deriving global vegetation index by season from AVHRR archived data is a precursor technology in this area.

Mineral and Oil Exploration

For many years the GEOSAT group has lobbied for space technology that supports the oil and mineral exploration industry. Several EOS sensors may be a step in this direction. HIRIS with its high spatial (30m) footprint and high spectral resolution shortwave infrared capability and pointable field-of-view (24km swath width) is well suited to potential mineral exploration capability. The Japanese ITIR will provide same orbit stereo observations at 15m resolution. This capability could benefit observation of geologic formations of interest to the oil industry. The MISR instrument provides a potential for stereo observations, but with a larger footprint over a wider swath. Again, the potential exists to develop commercial products tailored to a user in need of data in. poorly explored regions of the globe.

Products for the News Media

With multispectral capability and high resolution that is capable of observing interesting cultural features, the ITIR instrument may find potential users of its data to provide the nightly news or the printed media with color photography that might even show the deployment of opposing military confrontations, the extent of

Prod #	Cat	Parameter :: Qualifier	Type	Investigator	Instr.	Units	Accuracy Abs :: Rel	Temporal Resolution	Horizontal Resol. :: Domain	Vertical Resol. :: Domain	Time frame
2417	AR	Radiance :: Water_leaving	O :: Fl	Gordon et al	MODIS-N&T	mW/cm^2/sr/um	10% :: 5%	1 dy,wk,mo	20 km :: Ocean/GR	N/A :: Sfc	AL
2577	BM	Coccolith :: Detached	O :: Fl	Gordon,Clark	MODIS-N&T	mg-CaCO3/m^3	30% :: 10%	1 dy,wk,mo	20 km :: Ocean/GR	N/A :: TOO	AL
2578	BM	Coccolith :: Detached	O :: Fl	Gordon,Clark	MODIS-N&T	mg-CaCO3/m^3	30% :: 10%	1 dy,wk,mo	1 km :: Ocean/L	N/A :: TOO	AL
2591	BM	Pigment ::	O :: Fl	Gordon,Clark	MODIS-N&T	mg/m^3	30% :: 10%	1 dy,wk,mo	20 km :: Ocean/GR	N/A :: TOO	AL
2592	BM	Pigment ::	O :: Fl	Gordon,Clark	MODIS-N&T	mg/m^3	30% :: 10%	1 dy,wk,mo	1 km :: Ocean/RL	N/A :: TOO	AL
3199	OR	Attenuation_coef :: @490nm	O :: Fl	Gordon,Clark	MODIS-N&T	/m	25% :: 10%	1 dy,wk,mo	20 km :: Ocean/L, RL	N/A :: TOO	AL
3200	OR	Attenuation_coef :: @490nm	O :: Fl	Gordon,Clark	MODIS-N&T	/m	25% :: 10%	1 dy,wk,mo	1 km :: Ocean/L, RL	N/A :: TOO	AL
2573	BM	Chlorophyll :: Fluor_line_curv	O :: Fl	Hoge	MODIS-N	mW/cm^2/sr/um	25% :: 8%	1 dy,wk	1 km :: Ocean/R	N/A :: TOO	AL
2574	BM	Chlorophyll :: Fluor_line_curv	O :: Fl	Hoge	MODIS-N	mW/cm^2/sr/um	25% :: 8%	1 dy,wk	20 km :: Ocean	N/A :: TOO	AL
2593	BM	Pigment :: Curv	O :: Fl	Hoge	MODIS-N&T	mg/m^3	35% :: 15%	1 dy, 1 wk	1 km :: Ocean/R	N/A :: TOO	FL
2594	BM	Pigment :: Curv	O :: Fl	Hoge	MODIS-N&T	mg/m^3	35% :: 15%	1 dy, 1 wk	20 km :: Ocean/R	N/A :: TOO	FL
2585	BM	Phycoerythrin ::	O :: Fl	Hoge	MODIS-T	mg/m^3	200% :: 50%	1 dy,wk,mo	1 km :: Ocean/RL	N/A :: TOO	AL
2586	BM	Phycoerythrin ::	O :: Fl	Hoge	MODIS-T	mg/m^3	200% :: 50%	1 dy,wk,mo	20 km :: Ocean	N/A :: TOO	AL
2537	AR	Temperature_diff :: Day-night	O :: Fl	Huete	MODIS-N	K	1K :: 1K	1 dy	856 m :: Regional	N/A :: Sfc	FL
2847	AR	Brightness_index :: Soil	O :: Fl	Huete	MODIS-N&T	%	5% :: 5%	mo	1 km :: Land/R	N/A :: Sfc	AL
2895	AR	Color_index :: Soil	O :: Fl	Huete	MODIS-N&T	class	10% :: 5%	mo	1 km :: Land/R	N/A :: Sfc	AL
2896	AR	Mixture_model :: Spectral-Spatial	O :: Fl	Huete	MODIS-N&T	dimensionless	5-10% :: 0.05	1 dy	Pixel size :: Global	N/A :: Sfc	FL
2720	BT	Vegetation_index :: SoilABRDF_adjusted	O :: Fl	Huete	MODIS-N&T	dimensionless	[illegible]	1 dy,wk,mo	1 km :: Land/R	N/A :: Sfc	FL
2740	BT	Vegetation_index :: Soil_adjusted	O :: Fl	Huete	MODIS-N&T	dimensionless	[illegible]	dy,wk,mo	1 km :: Land/R	N/A :: Sfc	AL
3904	XX	XXX :: Data_char	O :: Fl	Justice,Strahler	MODIS-N&T	dimensionless	30,10, 5% ::	1 dy	1 km :: Global	N/A :: Sfc	FL
3905	XX	XXX :: Data_char	O :: Fl	Justice,Strahler	MODIS-N&T	dimensionless	30,10, 5% ::	1 dy	10 km :: Global	N/A :: Sfc	FL
3906	XX	XXX :: Data_char	O :: Fl	Justice,Strahler	MODIS-N&T	dimensionless	30,10, 5% ::	1 dy	50 km :: Global	N/A :: Sfc	FL
2609	BT	Duration :: Growing_season	O :: Fl	Justice	MODIS-N&T	dy	10 dy ::	1 yr	1 km :: Land	N/A :: Sfc	FL
2610	BT	Duration :: Growing_season	O :: Fl	Justice	MODIS-N&T	dy	10 dy ::	1 yr	10 km :: Land	N/A :: Sfc	FL
2749	BT	Vegetation_index ::	O :: Fl	Justice,Huete,et al	MODIS-N&T	dimensionless	0.01 :: 0.01	dy,wk,mo	10 km :: Land	N/A :: Sfc	AL
2750	BT	Vegetation_index ::	O :: Fl	Justice,Huete,et al	MODIS-N&T	dimensionless	0.01 :: 0.01	dy,wk,mo	0.5 km :: Land/R	N/A :: Sfc	AL
2751	BT	Vegetation_index ::	O :: Fl	Justice,Huete,et al	MODIS-N&T	dimensionless	0.01 :: 0.01	dy,wk,mo	1 km :: Land/R	N/A :: Sfc	AL
2259	AR	Heterogeneity :: Spatial	O :: Fl	Justice,Strahler	MODIS-N	W/m^2/sr/um	50% ::	1 dy	428 m :: Global	N/A :: Sfc	AL
2260	AR	Heterogeneity :: Spatial	O :: Fl	Justice,Strahler	MODIS-N	W/m^2/sr/um	20% ::	1 dy	856 m :: Global	N/A :: Sfc	AL
2262	AR	Heterogeneity :: Spatial	O :: Fl	Justice,Strahler	MODIS-T	W/m^2/sr/um	30% ::	1 dy	1.1 km :: Global	N/A :: Sfc	AL
2040	AR	Cloud_cover :: Areal_extent	O :: Fl	Kaufman	MODIS-N	km^2		mo	1 dg :: Global	N/A :: Sfc	FL
2092	AR	Cloud_cover :: Polygon	O :: Fl	Kaufman	MODIS-N	km		mo	1 dg :: Global	N/A :: Sfc	FL
2029	AR	Reflectance :: Directional	O :: Fl	Kaufman et al	MODIS-N	dimensionless	0.01 :: 0.005	1 dy	856 m :: Global	N/A :: Sfc	AL
2030	AR	Reflectance :: Directional	O :: Fl	Kaufman et al	MODIS-N	dimensionless	0.01 :: 0.005	1 dy	428 m :: Global	N/A :: Sfc	AL
2031	AR	Reflectance :: Directional	O :: Fl	Kaufman et al	MODIS-N	dimensionless	0.01 :: 0.005	1 dy	214 m :: Global	N/A :: Sfc	AL
2671	AR	Temperature :: Fire	O :: Fl	Kaufman, Justice	MODIS-N	C (K)	10 C (K) :: 5 C (K)	1 dy,wk	1 km :: Land/R	N/A :: Sfc	AL
2663	BT	Fire :: Count	O :: Fl	Kaufman, Justice	MODIS-N			1 dy,wk	1 km :: Land/R	N/A :: Sfc	AL
2664	BT	Fire :: Count	O :: Fl	Kaufman, Justice	MODIS-N			1 dy,wk	10 km :: Land	N/A :: Sfc	AL
2665	BT	Fire :: [illegible]	O :: Fl	Kaufman, Justice	MODIS-N			1 dy,wk	1 km :: Land/R	N/A :: Sfc	AL
2666	BT	Fire :: [illegible]	O :: Fl	Kaufman, Justice	MODIS-N			1 dy,wk	10 km :: Land	N/A :: Sfc	AL
2711	AR	Temperature :: Fire	O :: Fl	Kaufman, Justice	MODIS-N	C (K)	10 C (K) :: 5 C (K)	1 dy,wk	10 km :: Land	N/A :: Sfc	AL
1017	AC	Aerosol_mass ::	O :: Fl	Kaufman,Tanre	MODIS-N	g/m^2	30% :: 10%	dy,mo	0.5 dg :: Global/R	N/A :: Atmos	AL
1074	AH	H2O :: Integrated	O :: Fl	Kaufman,Tanre	MODIS-N	mm	5 mm :: 3 mm	1 dy	5 km :: Global	N/A :: Atmos	AL
2088	AR	Optical_depth :: Aerosol	O :: Fl	Kaufman,Tanre	MODIS-N	dimensionless	0.1 :: 0.05	dy,mo	0.5 dg :: Land	N/A :: Atmos	AL
2279	AR	Radiance :: Land_leaving	O :: Fl	Kaufman,Tanre	MODIS-N	W/m^2/sr/um	10% :: 5%	1 dy	1 km :: Land/R	N/A :: Sfc	AL
2280	AR	Radiance :: Land_leaving	O :: Fl	Kaufman,Tanre	MODIS-N	W/m^2/sr/um	10% :: 5%	1 dy	10 km :: Land	N/A :: Sfc	AL
2281	AR	Radiance :: Land_leaving	O :: Fl	Kaufman,Tanre	MODIS-N	W/m^2/sr/um	10% :: 5%	1 dy	0.5 km :: Land/R	N/A :: Sfc	AL
1764	AH	Cloud_drop_phase ::	O :: Fl	King	MODIS-N	[illegible]	90% Conf :: 90% Conf	1/dy	5 km :: Global	N/A :: Cloud	AL
1765	AH	Cloud_drop_phase ::	O :: Fl	King	MODIS-N	[illegible]	90% Conf :: 90% Conf	1 dy,mo	1 dg :: Global	N/A :: Cloud	AL
2081	AR	Cloud_cover :: Fraction	O :: Fl	King	MODIS-N	%	10% :: 5%	[illegible]	5 km :: Global	N/A :: Cloud	AL

TABLE 2: EXAMPLE OF SELECTED DATA PRODUCTS FOR THE MODIS FACILITY INSTRUMENT

earthquake damage, the details of an ash cloud from a volcanic explosion, or the acreage destroyed by fires in the tropical rain forest. Of course, mother nature has to cooperate with reasonable clear observing conditions, but given the opportunity the potential is there to service the news organizations.

Land Management

Geographic information systems are beginning to be a practical tool for local, regional, and national governments to use in land management. Documentation on utility and transportation infrastructure and zoning are already being incorporated. Remote sensing products will be increasingly beneficial to the monitoring of land use, of changes with time, and of natural resources. Several EOS instruments can benefit customers concerned with these issues. The value of the data from a given instrument depends on the scale of the observation, the field-of-view, and the temporal coverage. Third world customers could be a significant user of these kinds of data sets if the benefit to price ratio is within the scope of their resources. Value added services will have to work with these customers to develop the data needs of the land managers.

Ocean Fishing Fleet Harvest Forecasts

The MODIS instrument is specifically tailored to extend the present Coastal Zone Color Scanner capability to a higher precision product. Knowledge of seasonal trends in bioproductivity and potentially weekly knowledge of fishing area conditions conducive or destructive to commercial fish populations could be products derived from EOS data. With a FAX machine in everybody's floating office the data could be provided to the user in situ. The trick will be to develop enough experience to know what information is useful to be able to develop a commercial information product from the EOS data sets.

SUMMARY

Combinations of the data from selected instruments in the EOS payload has the potential to meet the needs of public and private customers. The development of commercial data products from these data requires value added services. The people that provide these value added services will have to understand the needs of their customers, the nature of the EOS data sets, and the proper combination of EOS data to meet the needs of their customers. Most of the initiative and imagination needed to create these products will come from the private sector. Data policy and pricing are issues to be worked over the next half decade.

SESSION D - HUMAN FACTORS ENGINEERING AND LIFE SCIENCES (PART 2)

Wednesday November 28, 1990

- Simulation Of Blood Flow Through An Artificial Heart
- Three-Dimensional Structure Of Human Serum Albumin
- A Noninvasive Measure Of Minerals And Electrolytes In Tissue
- Oxygen Production Using Solid-State Zirconia Electrolyte Technology
- Monitoring And Control Technologies For Bioregenerative Life Support Systems

PRECEDING PAGE BLANK NOT FILMED

N91

24056

UNCLAS

N91-24056

SIMULATION OF BLOOD FLOW THROUGH AN ARTIFICIAL HEART

Cetin Kiris, I-Dee Chang
Stanford University
Stanford, California

Stuart E. Rogers, Dochan Kwak
NASA Ames Research Center
Moffett Field, California

Abstract

A numerical simulation of the incompressible viscous flow through a prosthetic tilting disk heart valve is presented in order to demonstrate the current capability to model unsteady flows with moving boundaries. Both steady-state and unsteady flow calculations are done by solving the incompressible Navier-Stokes equations in three-dimensional generalized curvilinear coordinates. In order to handle the moving boundary problems, the chimera grid embedding scheme which decomposes a complex computational domain into several simple subdomains is used. An algebraic turbulence model for internal flows is incorporated to reach the physiological values of Reynolds number. Good agreement is obtained between the numerical results and experimental measurements. It is found that the tilting disk valve causes large regions of separated flow, and regions of high shear.

Introduction

Various types of prosthetic heart valves have been used widely as the replacements of natural valves since the first successful valve replacement performed in 1960. However, each of the valve design has some difficulties, which cause the artificial heart valve to be less efficient than the natural one. The difficulties which are related to the nonphysiological flow characteristics of the currently used prosthetic heart valves are: 1) Large pressure losses across the valves prevent the heart working efficiently; 2) Separated and secondary flow regions cause clotting; 3) High turbulent shear stress can damage the red blood cells. Having detailed knowledge of the flow quantities can help a design engineer improve the valve geometry, where a smooth flow is desired. Certain experimental studies[1-3] have pointed out the effects of the stagnation and recirculation regions and compared commonly used valve geometries. Since the experimental measurements provide flow characteristics for only certain regions of the flowfield, the numerical simulation of the flow through the artificial heart valve will be extremely helpful in the design and development stage of the prostheses.

Most of the numerical studies modeled the flow through the heart valve devices by excluding the moving boundary problems. Underwood and Mueller[4] obtained the flow characteristics for Kay-Shiley disk type valve using the stream function-vorticity formulation. Their results showed agreement with experimental data up to Reynolds number of 600. Idelsohn, Costa, and Ponso[5] modeled the flow through Kay-Shiley caged disk, Starr-Edwars caged ball, and Bjork-Shiley tilting disk valves and compared their performance. A maximum Reynolds number of 1500 was reached in their numerical study. In the above studies, the caged disk and caged ball geometries are axisymetric, and the tilting disk geometry is two-dimensional. In actual case, the tilting disk geometry is three-dimensional, the flow through heart valves is unsteady, and Reynolds numbers are as high as 6500. Peskin and McQueen[6] modeled the prosthetic heart valves in the numerical simulation of the flow in the heart. They used boundary forces derived from the energy function in order to model valve opening and closing, and they also modeled the elastic behavior of the walls. Their solution is obtained for the low Reynolds numbers in two dimensions using square cartesian mesh. McCracken and Peskin[7] applied a combination vortex-grid method for the blood flow through the mitral valve in two dimensions. This method is applied to the problems in which the solution does not have strong dependence on the Reynolds number. Peskin and McQueen[8] demonstrates the capability of modeling elastic behavior of the heart muscle by applying their extended three dimensional solution procedure to a toroidal tube. In order to obtain a solution procedure aimed at design improvements in prosthetic heart valves, the computation of steady-state and unsteady flow through the Bjork-Shiley tilting disk valve in three-dimensional configuration with the use of a grid embedding scheme is proposed in the current work. The equations are solved a in curvilinear generalized coordinate system, and the valve opening and closing are simulated by calculating the forces acting on the valve.

One of the biggest difficulties in the simulation of the flows in complicated three-dimensional configurations is the discretization of the physical domain with a single grid. The problem becomes more severe if one body in the domain of interest moves relative to another one as is seen in the tilting disk configuration. The use of a zonal approach[9] would be a practical solution of the moving boundary problem if the grids could be constrained to common boundaries. The chimera grid embedding technique[10,11] provides a greater flexibility for the grid motion. Instead of using common boundaries between grids, common regions are used. In the present work, the chimera approach is used to discretize the geometry of the disk valve. In addition, the procedure obtained for the heart valve configuration can be easily utilized for other unsteady incompressible viscous flows with moving boundaries, e.g., flow through Space Shuttle External Tank/Orbiter propellant feed line disconnect flapper valves.

In the first section, the method of solving the incompressible Navier-Stokes equations is described, and the algebraic turbulence model is summarized. Next, the geometry and the use of the chimera scheme are discussed. Following that is a presentation of the computed results obtained from the current approach.

Governing Equations and Method of Solution

The algorithm used in both steady-state and unsteady flow calculations is based on the method of artificial compressibility, which produces a hyperbolic system of equations by introducing a time derivative pressure term into the continuity equation.

The resulting incompressible Navier-Stokes equations can be written in a generalized curvilinear coordinate system (ξ, η, ζ) as follows

$$\frac{\partial Q}{\partial \tau} + \frac{\partial}{\partial \xi}(E - E_v) + \frac{\partial}{\partial \eta}(F - F_v) + \frac{\partial}{\partial \zeta}(G - G_v) = 0 \tag{1}$$

where Q, and convective flux vectors E, F, G are

$$Q = \frac{1}{J}\begin{bmatrix} p \\ u \\ v \\ w \end{bmatrix} \qquad E = \frac{1}{J}\begin{bmatrix} \beta U \\ \xi_x p + uU + \xi_t u \\ \xi_y p + vU + \xi_t v \\ \xi_z p + wU + \xi_t w \end{bmatrix}$$

$$F = \frac{1}{J}\begin{bmatrix} \beta V \\ \eta_x p + uV + \eta_t u \\ \eta_y p + vV + \eta_t v \\ \eta_z p + wV + \eta_t w \end{bmatrix} \qquad G = \frac{1}{J}\begin{bmatrix} \beta W \\ \zeta_x p + uW + \zeta_t u \\ \zeta_y p + vW + \zeta_t v \\ \zeta_z p + wW + \zeta_t w \end{bmatrix}$$

Here J, β, p, u, v, and w denote the Jacobian of transformation, the pseudocompressibility coefficient, pressure, and velocity components, respectively. The contravariant velocity components U, V, and W are defined as

$$\begin{aligned} U &= \xi_t + \xi_x u + \xi_y v + \xi_z w \\ V &= \eta_t + \eta_x u + \eta_y v + \eta_z w \\ W &= \zeta_t + \zeta_x u + \zeta_y v + \zeta_z w \end{aligned}$$

For an orthogonal grid assumption, the viscous flux vectors E_v, F_v, and G_v are given by

$$E_v = \frac{1}{ReJ}\begin{bmatrix} ((\xi_x^2 + \xi_y^2 + \xi_z^2)u_\xi \\ ((\xi_x^2 + \xi_y^2 + \xi_z^2)v_\xi \\ ((\xi_x^2 + \xi_y^2 + \xi_z^2)w_\xi \end{bmatrix}$$

$$F_v = \frac{1}{ReJ}\begin{bmatrix} (\eta_x^2 + \eta_y^2 + \eta_z^2)u_\eta \\ (\eta_x^2 + \eta_y^2 + \eta_z^2)v_\eta \\ (\eta_x^2 + \eta_y^2 + \eta_z^2)w_\eta \end{bmatrix}$$

$$G_v = \frac{1}{ReJ}\begin{bmatrix} ((\zeta_x^2 + \zeta_y^2 + \zeta_z^2)u_\zeta \\ ((\zeta_x^2 + \zeta_y^2 + \zeta_z^2)v_\zeta \\ ((\zeta_x^2 + \zeta_y^2 + \zeta_z^2)w_\zeta \end{bmatrix}$$

where Re is the Reynolds number.

In the steady-state formulation, the time derivatives are differenced using the Euler backward formula. The equations are solved iteratively in pseudo-time until the the solution converges to a steady state. In the time-accurate formulation, the time derivatives in the momentum equations are differenced using a second-order, three-point, backward-difference formula. The equations are iterated to convergence in pseudotime for each physical time step until a divergence-free velocity field is obtained. The numerical method uses a second-order central difference for viscous terms and a higher order flux-difference splitting for the convective terms. The ξ derivative of the convective flux E can be written as

$$\frac{\partial E}{\partial \xi} \approx \frac{[E_{i+1/2} - E_{i-1/2}]}{\Delta \xi}$$

The numerical flux $E_{i+1/2}$ is defined as follow

$$E_{i+1/2} = \frac{1}{2}\left[E(Q_{i+1}) + E(Q_i) - \phi_{i+1/2}\right] \tag{2}$$

where the $\phi_{i+1/2}$ is a dissipation term. The order of the scheme is determined by the definition of the dissipation term $\phi_{i+1/2}$. For $\phi_{i+1/2} = 0$, the differencing is reduced to a second-order central difference scheme. A first-order upwind flux is defined by

$$\phi_{i+1/2} = \left(\Delta E^{+}_{i+1/2} - \Delta E^{-}_{i+1/2}\right) \tag{3}$$

and a third-order upwind flux is given by

$$\phi_{i+1/2} = -\frac{1}{3}\left(\Delta E^{+}_{i-1/2} - \Delta E^{+}_{i+1/2} + \Delta E^{-}_{i+1/2} - \Delta E^{-}_{i+3/2}\right) \tag{4}$$

where $\Delta E^{\pm}$ is the flux difference across positive or negative traveling waves, and is computed as

$$\Delta E^{\pm}_{i+1/2} = A^{\pm}(\bar{Q})\Delta Q_{i+1/2}$$

here $A^{\pm}$ is the plus (minus) Jacobian matrix. The Δ operator, and $\bar{Q}$ are given by

$$\Delta Q_{i+1/2} = Q_{i+1} - Q_i$$

$$\bar{Q} = \frac{1}{2}(Q_{i+1} + Q_i)$$

An implicit delta law form approximation to Eq.(1) after linearization in time and the use of approximate Jacobians of the flux differences results in a 4 x 4 block heptadiagonal matrix equation. The matrix equation is solved iteratively by using a nonfactored line relaxation scheme, which maintains stability and allows a large pseudotime step to be taken. At each sweep direction, a tridiagonal matrix is formed and off line terms of the matrix equation are moved on the right-hand side of the equation. Details of the numerical method are given in Refs. 12-14.

An algebraic mixing-length turbulence model, which is presented in Ref. 15, is utilized in the present computations. The turbulent eddy viscosity is taken as

$$\nu_t = l^2|w|$$

where l is the mixing-length, and $|w|$ is the absolute value of vorticity. In order to account the effect of more than one wall in the region close to the tilting disk, the mixing-length is given as

$$1/l = \left(\sum_{k=1}^{n} 1/(cy_k D_k)\right)$$

where n is the number of walls, $c = 0.4$ is the Von Karman's constant, y_k is the distance from the k^{th} surface, and D_k is the Van Driest damping factor for the k^{th} surface.

Geometry and Grid System

In the Bjork-Shiley tilting disk heart valve, the tilting disk is placed in front of the sinus region of the human aorta. The aortic root has three sinuses about 120 degree apart from one another. The tilting disk valve model used in this computation is simplified by assuming that the sinus region of aorta has a circular cross-section. The cage and struts which hold the free-floating disk inside the sewing ring are not included in the geometry. It is also assumed that the walls do not have an elastic deformation. The computational geometry used in unsteady flow computations is given in Figure 1. The channel length is taken to be five aorta diameters long. The disk motion is illustrated by showing three different positions of the disk. The disk angles shown are 75, 50, and 30 degrees as measured from the centerline of the aorta. The tilting disk is allowed to rotate about the horizontal axis that is 1/6 of a disk diameter below the center of the disk. Because of this asymmetric disk orientation, the flow is three dimensional.

The chimera grid embedding technique, which has been successfully used for external flow problems, has been employed by using two overlapped grids as shown in Figure 2-a. Grid 1 contains 17,199 points which are distributed 63 x 21 x 13 in the ξ, η, and ζ directions. It occupies the whole region in the aorta from entrance to exit, and remains stationary. Grid 2 has 4,725 points as distributed 25 x 21 x 9. It wraps around the tilting disk, and moves with the disk. The lateral symmetry planes of the two grids are shown in Figure 2-b in order to demonstrate how the grid embedding scheme is applied to the present problem. Grid points which lie within the disk geometry and outside the aorta grid are excluded from the solution process. These excluded points are called hole points, and the immediate neighbors of the hole points are called fringe points. The information is passed from one grid to another one via fringe and grid boundary points by interpolating the dependent variables. Tri-linear interpolation is used in the present computations. In order to distinguish the hole and fringe points from regular computational points, an IBLANK array is used in the flow solver. For hole, grid boundary, and fringe points IBLANK is set to zero, otherwise it is set to one. In order to exclude the hole and grid boundary points from the solution procedure at each time step, the ΔQ solution is processed as follows

$$Q^{n+1} = Q^n + \Delta Q \cdot IBLANK$$

Since 3rd order flux-differencing is used for convective terms, the order of differencing needs to be reduced to the second order differencing near the fringe points.

Computed Results

Presented here are the results of steady-state flow and unsteady flow with the disk motion in the configuration mentioned above. The problems are nondimensionalized by using the entrance diameter as a unit length, and the average inflow velocity as the unit velocity. The geometries used in the steady-state and unsteady calculations are similar to the geometries used in the experimental studies in Ref. 1,2 and Ref. 3 respectively. In order to reduce the computational effort and memory size, the inflow and outflow boundaries are placed a short distance from the region of interest in comparison with the boundaries in the experimental studies. In addition, the exact shape of the sinus region of aorta used in the experiments is not available at present. Because of these differences between experimental and computational configurations, there are small differences between experimental measurements and present computations.

At a viscous no-slip surface, both the velocity and the pressure gradient normal to the wall are specified to be zero. At the inflow boundary, the velocities are specified, and the pressure is determined from the characteristic boundary condition. At the outflow boundary, static pressure is specified, and velocities are calculated from the three characteristic waves traveling out of the computational domain.

Steady-state calculations for the 30 degree disk orientation have been carried out for Reynolds numbers in the range of 2000 to 6000, in which experimental data is available. The Reynolds number is based on the diameter and the mean velocity at the entrance of the channel. Figure 3 shows the convergence history for a Reynolds number of 5972. Both averaged residual and maximum divergence of velocity have dropped ten order of magnitude in 600 iterations. For an overlapped grid application, the convergence is shown to be very fast. The values for grid 1 are drawn with solid lines, and the values for grid 2 are drawn with dashed lines. The computing time per grid point per iteration is about $2x10^{-4}$ sec. Figure 4 shows the pressure drop across the Bjork-Shiley tilting disk valve at different flow rates of physiological interest. $P1$ and $P2$ are the pressures at the points located 150 mm, and 20 mm upstream from the disk at the centerline of the channel respectively. In order to compare numerical results with the experimental measurements given in Ref. 2, the numerical results are redimensionalized. The computed and measured axial velocity profiles at 42 mm downstream from the disk are shown in Figure 5. Figure 5-a demonstrates the horizontal plane, where axial velocity profiles are given, through the center of the channel. Axial velocity profiles given in Figures 5-b through 5-d illustrate how the stagnation region, which is created by the tilting disk valve, is dominant in the region of 1.5 disk diameter downstream. The numerical results are shown with dots and the experimental results are shown with triangles. The numerical results compare favorably with the experimental measurements given in Refs. 1-2.

Velocity vectors on the lateral symmetry plane are given in Figure 6 for a Reynolds number of 5972. The flow, which is directed to the upper part of the aorta, generates vortices in the sinus region of the aorta and a large separated region along the lower wall of the aorta. Since separated and low flow regions have potential for thrombus formation, clotting may occur on the upper sinus region and the lower wall of the aorta. The flow is highly accelerated near the tilting disk and the upper wall. Figure 7 shows vorticity magnitude contours on the surface of the channel, inflow surface, and outflow surface of the disk, respectively. It is assumed that maximum vorticity magnitudes indicate the regions of high shear. The sewing ring surface and the edges of the disk are the regions having maximum vorticity magnitude. The upper wall of the channel also has considerably high vorticity magnitudes.

Unsteady flow calculations have been carried out in order to demonstrate and analyze the flow during disk opening and closing. For the present computation, one cycle of valve opening and closing requires 70 physical time steps. During each time step, subiterations are carried out until maximum divergence of velocity and maximum residual drop below 10^{-3}. The computing time required for one cycle of the valve opening and closing is aproximately 5 hr. During the valve opening, inflow velocity is imposed at the entrance of the channel. The inflow velocity is chosen as a sinuous function in time. The forces acting on the disk are calculated, and the disk rotation angle is determined. For large disk rotation angle, some information may be lost between the grids when the grid embedding technique is used. In order to prevent the information loss, the maximum allowed disk rotation angle at each physical time step is taken to be less than three degrees. As soon as the disk reaches its fully opened position, which is 30 degrees measured from the horizontal plane, the flow direction is reversed by imposing the inflow velocity at the exit of the channel.

Figures 9-a through 9-f illustrate the velocity vectors on the lateral symmetry plane at t/T = 0.13, 0.285, 0.385, 0.53, 0.685, and 0.8 respectively. T is a period of one cycle during the valve opening and closing. The velocities are very high in the region between the disk and the channel wall as shown in Figure 9-a. During the disk opening, two vortices are formed at the upper and lower edges of the disk. The flow starts to separate behind the disk and reattaches to the wall as shown in Figure 9-b. The stagnation region behind the disk moves downstream as the disk rotates. Highly skewed velocity profiles are seen downstream from the disk as illustrated in Figure 9-c. The growth of the vortices has also been observed in the sinus region of the aorta while the flow opens the valve. Along the lower wall a separation region is formed. Figure 9-d shows the beginning of the valve closing. At this moment, the location of imposing the inflow velocity is moved from entrance to exit. Major flow near the upper wall of the channel forms a recirculation region downstream from the disk. With the help of this recirculation, the lower wall of the channel becomes the major flow region during the valve closing, and upper wall region becomes the low flow region.

The vorticity magnitude contours on the surface of the channel at t/T = 0.13, 0.285, and 0.385 are shown in Figures 8-a through 8-c in order to indicate the regions of high shear. At the beginning of valve opening, jet-like flow between the sewing ring and the disk cause very high vorticity magnitudes as shown in Figure 8-a. During the disk rotation, the high vorticity region on the upper wall of the channel moves from the sinus region to downstream as seen in Figures 8-a through 8-c. The results of more realistic flow calculations, such as the flow through the Pennsylvania State artificial heart including tilting disk valve opening and closing, will be reported in the future.

Concluding Remarks

The solution procedure for unsteady incompressible viscous flow computations has been extended with the incorporation of the grid embedding approach. This has been used to simulate both steady-state and unsteady flow through a tilting disk valve. The physiological values of the Reynolds number have been achieved with the use of a simple mixing-length turbulence model. The numerical results for 30 degree disk orientation were compared against the experimental data, and good agreement was obtained. The flow during the disk opening and closing were simulated within a reasonable computing time. The present capability of simulating complicated internal flow problems with moving boundaries is demonstrated. The procedure obtained here is quite general and applicable for various types of complicated geometries.

Acknowledgments

This work is partially supported by the NASA Technology Utilization office. Computer time is provided by the Applied Computational Fluids branch of NASA Ames Research Center.

References

[1] Yoganathan, A. P.,Concoran, W. H. and Harrison, E. C., "In Vitro Velocity Measurements in the Vicinity of Aortic Prostheses," *J. Biomechanics.* , Vol 12, pp. 135-152, 1979.

[2] Yoganathan, A. P.,Concoran, W. H. and Harrison, E. C., "Pressure Drops Across Prosthetic Aortic Heart Valves Under Steady and Pulsatile Flow," *J. Biomechanics.* , Vol 12, pp. 153-164, 1979.

[3] Figliola, R. S. and Mueller, T. J., "On the Hemolytic and Thrombogenic Potential Occluder Prosthetic Heart Valves from In-Vitro Measurements," *J. Biomech. Engng.* , Vol 103, pp. 83-90, 1981.

[4] Mueller, T. J., "Application of Numerical Methods in Physiological Flows," *Numerical Methods in Fluid Dynamics.* , 1978

[5] Idelsohn, S. R., Costa, L, E.. and , Ponso, R., "A Comparative Computational Study of Blood Flow Through Prosthetic Heart Valves Using the Finite Element Method," *J. Fluid Dynamics.* , Vol 18, No 2, pp. 97-115, 1985.

[6] Peskin, S. C., McQueen, D. M., "Modeling Prosthetic Heart Valves for Numerical Analysis of Blood Flow in the Heart," *J. Comp. Physics.* , Vol 37, pp. 113-132, 1980.

[7] McCracken, M. F., Peskin, S. C. "A Vortex Method for Blood Flow Through Heart Valves," *J. Comp. Physics.* , Vol 35, pp. 183-205, 1980.

[8] Peskin, S. C., McQueen, D. M., "A Three-Dimensional Computational Method for the Blood Flow in the Heart," *J. Comp. Physics*, Vol 81, pp. 372-405, 1989.

[9] Rai, M. M. "Unsteady Three-Dimensional Navier- Stokes Simulations of Turbine Rotor-Stator Interaction," AIAA Paper No. 87-2058, 1987.

[10] Dougherty, F.C., Benek, J. A., and Steger, J.L., "On Applications of Chimera Grid Schemes to Store Separation," NASA Techn. Memorandum 88193, 1985.

[11] Benek, J. A., Buning, P.G. and Steger, J.L., "A 3-D Chimera Grid Embedding Technique," AIAA Paper No. 85-1523, 1985.

[12] Rogers, S. E. and Kwak, D., "An Upwind Differencing Scheme For the Time-Accurate Incompressible Navier-Stokes Equations," AIAA Paper No. 88-2583, 1988.

[13] Rogers, S. E., Kwak, D. and Kiris, C., "Numerical Solution of the Incompressible Navier-Stokes Equations for Steady-State and Time-Dependent Problems," AIAA Paper No. 89-0463, 1989.

[14] Rogers, S. E., Kutler, P., Kwak, D. and Kiris,C, "Numerical Simulation of Flow Through an Artificial Heart." NASA Techn. Memorandum 102183, 1989.

[15] Burke, R. W. "Computation of Turbulent Incompressible Wing-Body Junction Flow," AIAA Paper No. 89-0279, 1989.

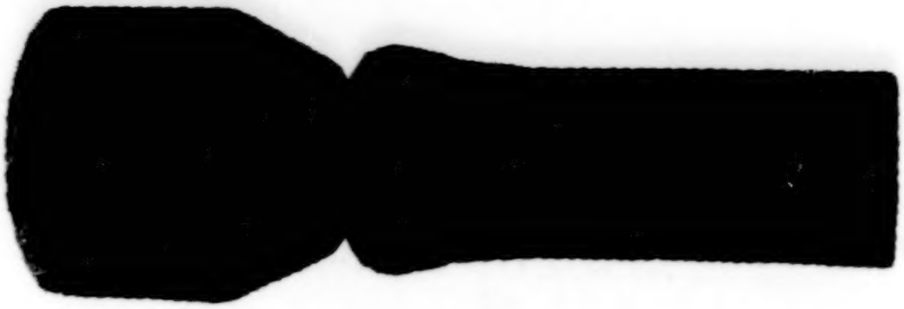


Figure 1. Tilting disk geometry showing valve opening.

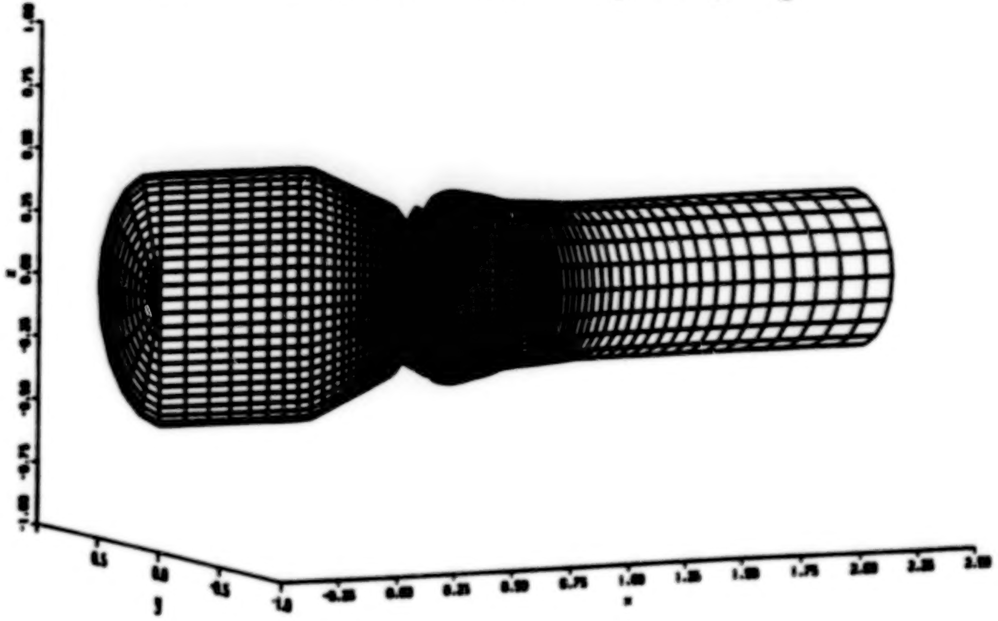

1.00
0.75
0.50
0.25
z
0.00
-0.25
-0.50
-0.75
-1.00
0.5
0.0
-0.5
-1.0
y
-0.25
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
x

Figure 2. a) Perspective view of two overlapped grids.

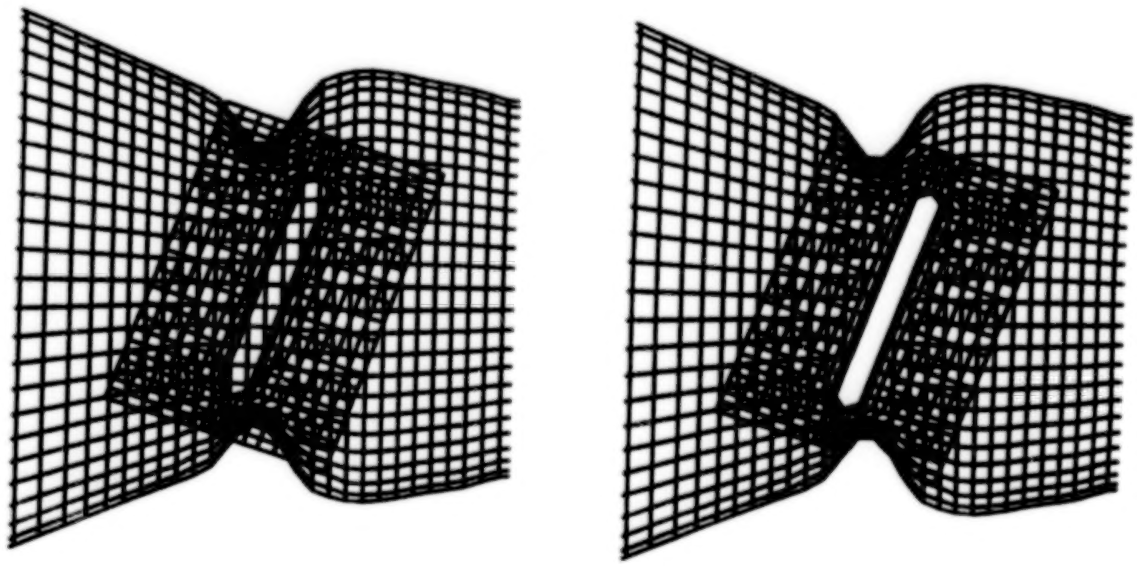


Figure 2. b) Side views of two overlapped grids before and after the hole points are excluded.

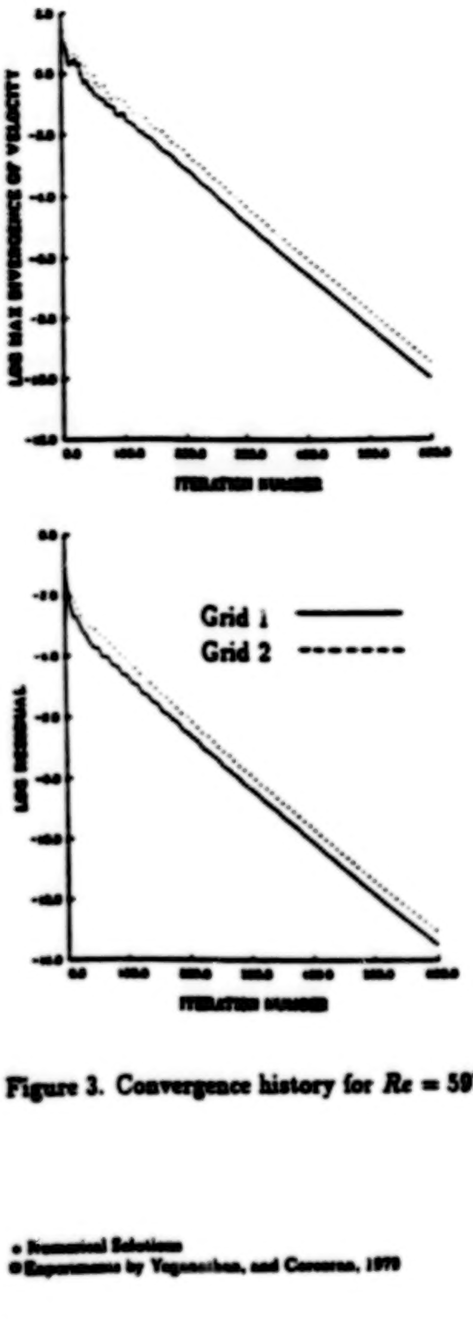


Figure 3. Convergence history for *Re* = 5972.

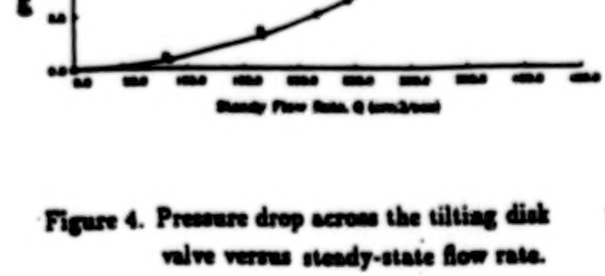
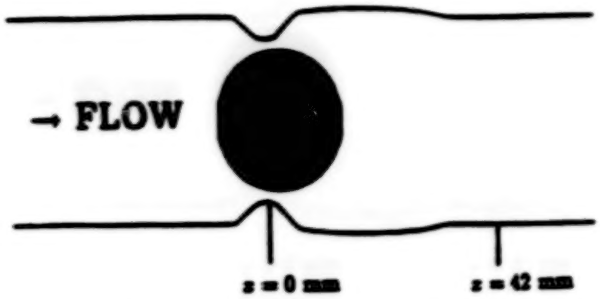
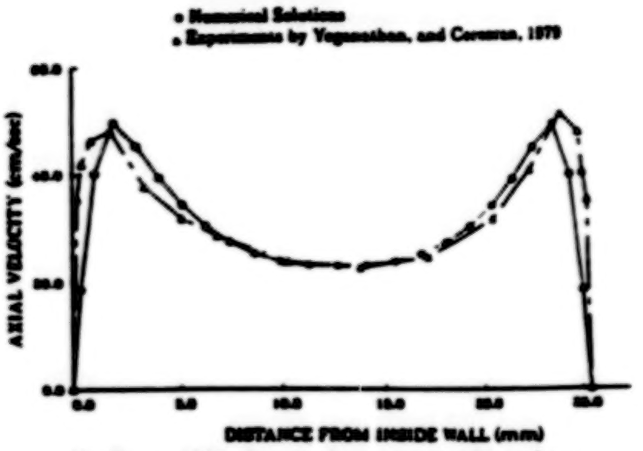


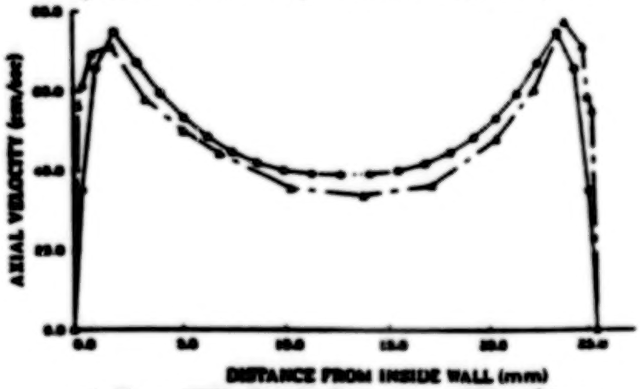
Figure 4. Pressure drop across the tilting disk valve versus steady-state flow rate.



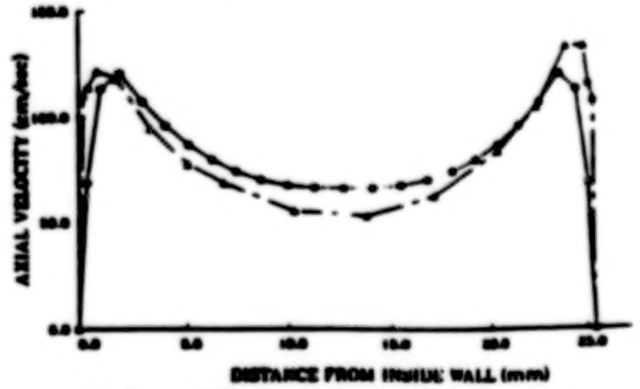
a) Top view of horizontal plane



b) *Re* = 2390, Steady flow rate = 167 cm^3/sec.



c) *Re* = 3580, Steady flow rate = 250 cm^3/sec.



d) *Re* = 5972, Steady flow rate = 417 cm^3/sec.

Figure 5. Top view of horizontal plane and axial velocity profiles on the horizontal plane for different flow rates.

ORIGINAL PAGE IS
OF POOR QUALITY

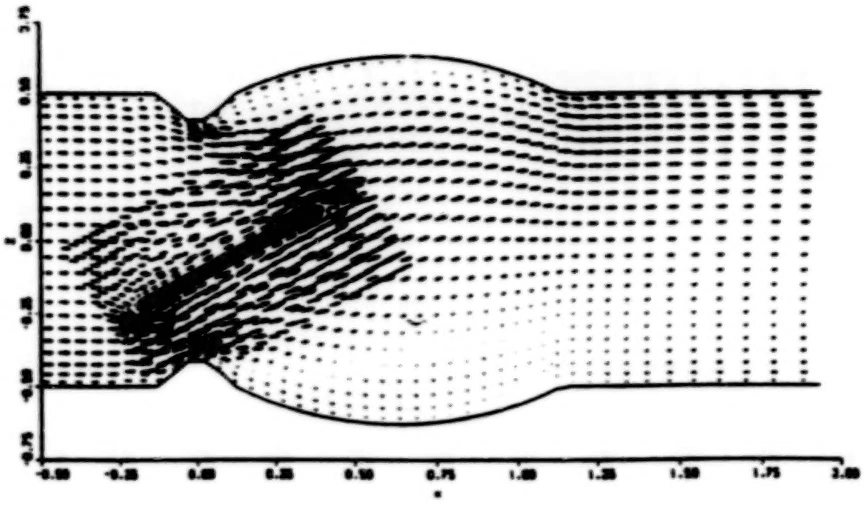
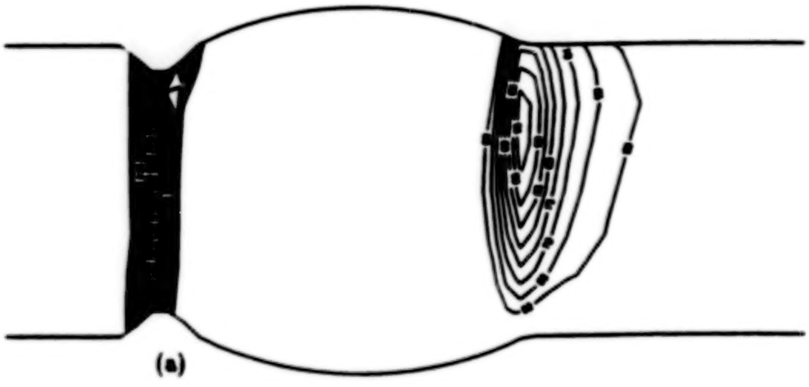


Figure 6. Side view of velocity vectors on the vertical plane.



(a)

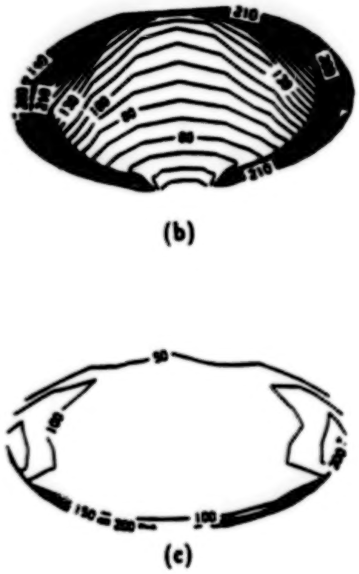
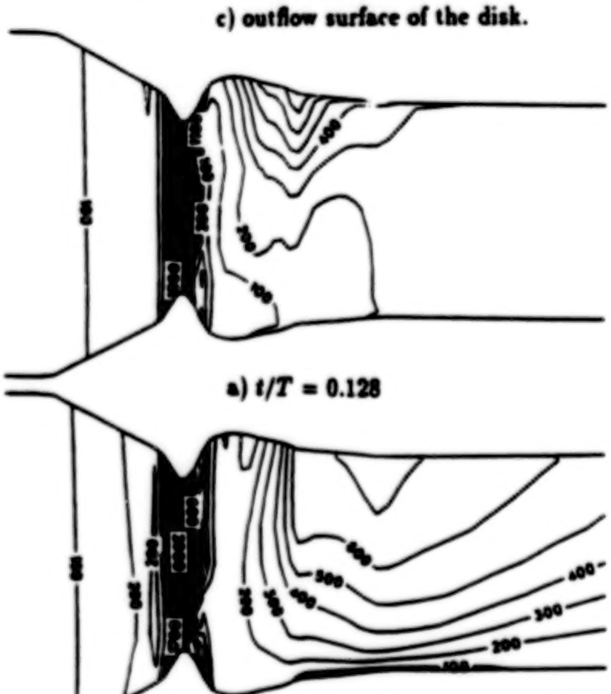
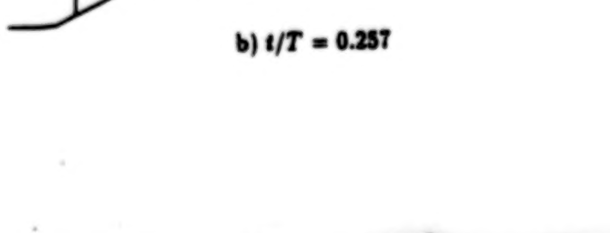


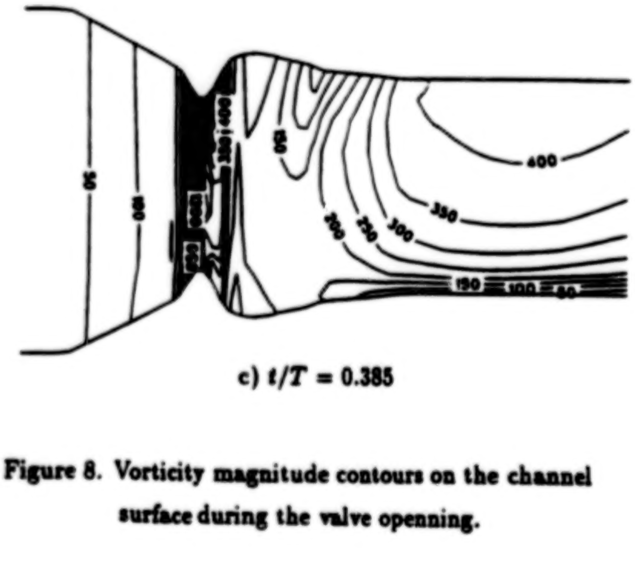
Figure 7. Vorticity magnitude contours on the a) channel surface. b) inflow surface of the disk, c) outflow surface of the disk.



a) $t/T = 0.128$

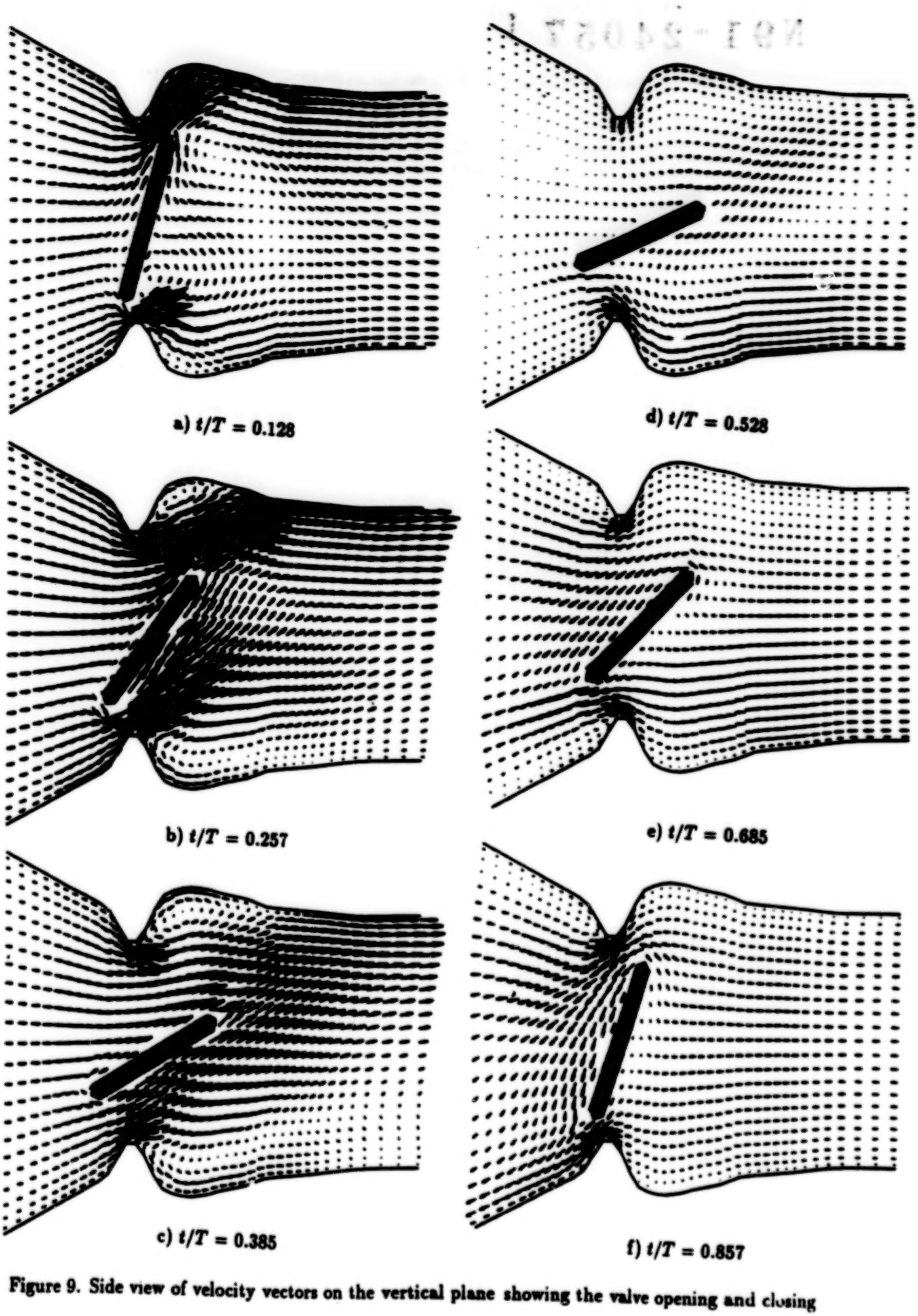


b) $t/T = 0.257$



c) $t/T = 0.385$

Figure 8. Vorticity magnitude contours on the channel surface during the valve openning.



a) $t/T = 0.128$

d) $t/T = 0.528$

b) $t/T = 0.257$

e) $t/T = 0.685$

c) $t/T = 0.385$

f) $t/T = 0.857$

Figure 9. Side view of velocity vectors on the vertical plane showing the valve opening and closing

N91

24057

UNCLAS

N91-24057

THREE-DIMENSIONAL STRUCTURE OF HUMAN SERUM ALBUMIN

**Daniel C. Carter, Xiao-min He, Pamela D. Twigg
and Elena Casale**

**The Space Science Laboratory
ES76 Biophysics
Marshall Space Flight Center, AL 35812**

All correspondence should be to Daniel C. Carter at the above address. Phone is (205) 544-5492. Fax is (205) 544-1777.

ABSTRACT

The binding locations to human serum albumin (HSA) of several drug molecules have been determined at low resolution using crystallographic methods. The principal binding sites are located within subdomains IIA and IIIA. Preliminary studies suggest that an approach to increasing the in vivo efficacy of drugs which are rendered less effective or ineffective by virtue of their interaction with HSA, would be the use of competitive displacement in drug therapies and/or the development of a general inhibitor to the site within subdomain IIIA. These findings also suggest that the facilitated transfer of various ligands across organ/circulatory interfaces such as the liver, kidney, and brain may be associated with binding to the IIIA subdomain.

Although serum albumin plays several important roles as the major protein of the circulatory system, it has been suggested that the molecule's high affinity for low molecular weight ligands may be its principal function (1). Over the past several decades there have been voluminous works describing the binding affinities and predicted locations of various endogenous and exogenous ligands to HSA. The reader is referred to the informative reviews by Kragh-Hannsen (2), Fehske et. al. (3), and Peters (4) and references cited therein. Previously, we described the structure of HSA as being comprised of three domains each of which were found to be the product of two helical subdomains consistent with the current understanding of the molecule's primary structure (5,6). These six subdomains have since been referred to as IA, IB, IIA, IIB, etc. We also reported that the binding of a chemically diverse group of compounds occurs primarily within two of these helical subdomains. These findings are in further agreement with several studies which indicate the presence of two to three high affinity binding sites within the molecule (2). More extensive experiments have recently been conducted by our group in an effort to identify the classical binding sites described by Sudlow and others (4,7). Sudlow grouped drug binding to HSA into three distinct categories, which he denoted as Site I for warfarin, Site II for the diazapines, and Site III for digitoxin. Although there is no complete agreement within the literature regarding these sites, warfarin is believed to bind primarily in domain II (3,4,8), the diazapines in domain III (9), and digitoxin, in an independent but unknown location (3,4). other studies indicate, however, that digitoxin competes with salicylates and warfarin for a common binding site (2,10). In this paper we present our preliminary findings regarding the major binding locations on HSA for several pharmaceutical compounds.

At the current resolution of 4.0 A the primary binding pockets within IIA and IIIA are cavities of approximately 8 A X 12 A X 18 A which are exposed to the solvent channels through an opening of -9-10 A in diameter. Large, well-ordered side chains protrude into these binding pockets where they appear to play a major role in ligand binding through stacking interactions. The identity of these amino acid residues should become evident once the model is fit to the electron density at higher resolution.

The general binding locations of several exogenous and endogenous ligands have been determined at various low resolutions using standard crystallographic methods, table 1. An illustration of various binding locations on HSA for several ligands is shown pictorially in figure 1. There exists, based on the current binding studies, at least four distinct major binding regions on HSA. Location 1, subdomain IIIA, which is the most active and accommodating on HSA with few exceptions, e.g. warfarin, displays high affinities for most organic compounds. Location 2, subdomain IIA, has similar binding properties to IIIA but in these preliminary studies shows an absence of long-chain fatty acid binding. Based on the amino acids implicated, IIA may be inferred to be the bilirubin high affinity site. Location 3, the IA-IB interface is an additional binding site for aspirin (although weak), decanol, and possibly the short chain fatty acids. Location 4, subdomain IB, possesses a high affinity surface, thus far unique to long chain fatty acids and tryptophan. The general binding locations of the compounds of Sudlow's group, warfarin (Site I) , diazapam (Site II), and digitoxin (Site III), were determined crystallographically. The crystallographic studies are in general agreement with the expected locations indicating major binding sites for warfarin within IIA, and for diazapam in IIIA. Digitoxin, also shows a major binding site within subdomain IIIA and a secondary site at the interface between IA and IIA. Sudlows Site III then, differs from Site II only in the nature of the secondary binding sites.

It is interesting to note that the amino acid sequences most highly conserved between bovine, human , and rat albumins; Cys-34, Trp-212, sequence 143 through 155 and sequence 244 through 263 (11), all appear to be in regions involved in ligand binding. For example, residues 143 through 155, an aromatic sequence, corresponds with the region of Location 4, the second major long-chain fatty acid binding site. Certainly the correlation of the conserved amino acids sequences with occurrence of binding sites within HSA supports the idea that the principal role of this molecule in the circulatory system is to aid in the transport, distribution, and metabolism of both essential and foreign small molecules. Given the unique structure of serum albumin and the following interesting observations: (i) there is an asymmetric charge distribution within the primary structure of HSA at pH 7, of -9, -8, and -2 for domains I, II, and III respectively, thus indicating a potential amphoteric nature, (ii) HSA has been shown to interact with the cell membrane thereby aiding in the delivery of certain molecules, e.g. thyroxine, to the cell (12), (iii) subdomain IIIA appears to be the primary binding site for a variety of molecules, (iv) serum albumin is implicated in the facilitated transfer of ligands across various organ/circulatory interfaces, such as the liver, intestine , kidney, and brain; it is tempting to speculate whether the serum albumins possess membrane spanning properties. If this were the case, one might expect that HSA would unfold at the hinge region between IIA and IIB allowing the more lipophylic tail (subdomains IIB, IIIA, and IIIB) of HSA to span the membrane while the more polar head of the molecule (subdomains IA, IB, IIA) would remain on the surface of the membrane, thus releasing the bound ligands in subdomain IIIA into the cell. Perhaps then, ligand binding within subdomain IIIA is a prerequisite for the facilitated transfer of many endogenous and exogenous compounds at organ/circulatory interfaces.

As we continue to characterize the nature of bound ligands to serum albumin and correlate this information with a wealth of drug and ligand binding data in the literature; it will become possible to predict important displacement interactions for a variety of endogenous and exogenous ligands. Accordingly, this suggests that an approach to increasing the in vivo efficacy of drugs which are rendered less effective or ineffective by virtue of their interaction with HSA would be the use of competitive displacement in drug therapies and/or the development of a general inhibitor to the site within subdomain IIIA.

ACKNOWLEDGEMENTS

This research was supported by the Office of Space Science and Applications of the National Aeronautics and Space Administration (RTOP 674-23-08-17). X-M. H., P. D. T., and E. C. were supported under a contract with the Universities Space Research Association. We thank D. Donovan and Glenn Roberts of Roberts Associates for helpful assistance.

REFERENCES

1 Putnam, F. W.: The Plasma Proteins, 2nd ed., vol. 4. Academic. London.

2. Kragh-Hannsen, U.: Molecular aspects of ligand binding to serum albumin. Pharmacol. Rev. vol 33, 1981, pp 17-53.

3. Fehske, K., J., Muller, W. E., and Wollert, U.: The location of drug binding sites in human serum

albumin. Biochem. Pharmacol. vol 30, 1981, pp 687-692.

4. Peters, T.: Serum albumin. Adv. Protein Chem. vol 37, 1985, pp 161-245.

5. Carter, D. C. , He, X. M. , Munson, S. H. , Twigg, P. D. , Broom, M. B., and Miller, T. Y.: Three-dimensional structure of human serum albumin. Science vol 244, 1989, pp 1195-1198.

6. Carter, D. C. and He, X-M.: Structure of Human serum Albumin. Science vol 249, 1990, pp 302-303.

7. Sudlow, G., Birkett, D. J., and Wade, D. N.: Further characterization of specific drug binding sites on human serum albumin. Mol. Pharmacol. 12:1052-1061, 1976.

8. Sjoholm, I. and Ljungstedt, I.: Studies on the tryptophan and drug binding properties of human serum albumin fragments by affinity chromatography and circular dichroism measurements. J. Biol. Chem. vol 248, 1973, pp 8434-8441.

9. Bos, O. J. M., Fischer, M. J. E., Wilting, J., and Janssen, L. H. M.: Drug-binding and other physicochemical properties of a large tryptic and a large peptic fragment of human serum albumin. Biochem. Biophysica Acta vol 953, 1988, pp 37-47.

10. Perrin, J. H., and Nelson, D. A.: Competitive binding of two drugs for a single binding site on albumin: a circular dichroic study. J. Pharm. Pharmacol. vol 25, 1973, pp 125-130.

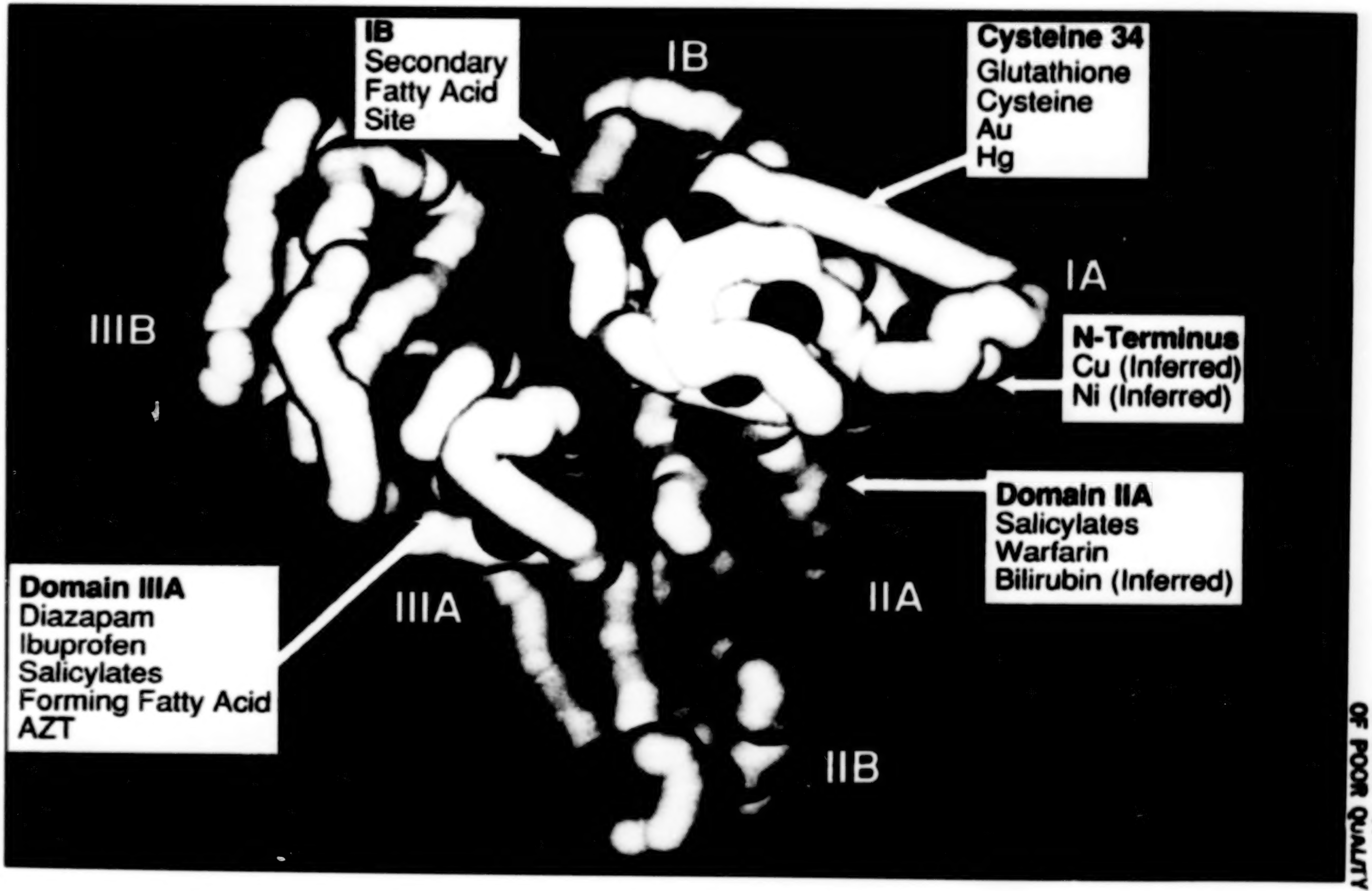
11. Morinaga, T., Sakai, M., Wegmann, T. G., and Tamaoki, T.: Primary structures of human alpha-fetoprotein and its MRNA. Proc. Natl. Acad. Sci. U.S.A. vol 80, 1983, pp 4604-4608.

12. Pardridge, William, M.: Plasma protein-mediated transport of steroid and thyroid hormones. Am. J. Physiol. vol 252, 1987, (Endocrinol. Metab. 15): E157-E164.

TABLE I LIGAND BINDING LOCATIONS TO HSA

Data Set	Ligand	D	N	Rf	Obsv. Loc.	Pred. Loc.
HTRP1A	Tryptophan	5.0	3198	0.093	IIIA,IB, IIA^IB	
HPAL1A	Palmitate	5.9	2432	0.066	IIIA,IB	
HMYR1A	Myristate	6.0	1923	0.095	IIIA	
HMRGL1A	Glucose	5.8	3106	0.10	IIIA	
HSAA1A	Aspirin	4.0	7362	0.11	IIIA,IIA	
HMRWN1A	Warfarin	5.0	2555	0.167	IIA	
HSAV1A	Diazapam	6.8	2075	0.118	IIIA	
HDGX1A	Digitoxin	5.0	3751	0.137	IIIA,IA^IIA	
HCHLO1A	Chlofibrate	6.0	2175	0.138	IIIA	
HMRIB1A	Ibuprofen	6.0	2402	0.215	IIIA	
HAZT1A	AZT	4.0	7548	0.124	IIIA	

It should be noted that these data represent preliminary research into ligand binding of HSA. The binding results can also be affected by the composition of the solvent, the pH, and protein packing within the crystal. Where D = minimum d-spacing for data collected, N = number of unique reflections measured, Rf = agreement between native and ligand for Fobs.s



IB
Secondary
Fatty Acid
Site

Cysteine 34
Glutathione
Cysteine
Au
Hg

N-Terminus
Cu (Inferred)
Ni (Inferred)

Domain IIA
Salicylates
Warfarin
Bilirubin (Inferred)

Domain IIIA
Diazapam
Ibuprofen
Salicylates
Forming Fatty Acid
AZT

FIGURE 1

ORIGINAL PAGE IS
OF POOR QUALITY

N91

24058

UNCLAS

N91-24058

A NON-INVASIVE MEASURE OF MINERALS AND ELECTROLYTES IN TISSURE

Sara B. Arnaud, M.D.
Research Scientist
Ames Research Center

Burton Silver, Ph.D.
Chairman of the Board, Director of Research
Intracellular Diagnostics, Inc.

INTRODUCTION

In clinical practice, metabolic disturbances in patients are monitored routinely by measurements of electrolytes and minerals in blood and urine. This method was used during the Skylab missions to detect potential metabolic problems in astronauts. Blood samples obtained before, during, and after the 28, 59 and 84 day flights revealed alterations in the concentration of calcium, phosphorus, sodium and potassium (1). Although these findings did not seem to present a problem to the space travelers, further monitoring is indicated because deviations from pre-flight probably represent part of man's adaptation to microgravity, a poorly understood process. A means for monitoring changes in electrolytes and minerals without venipuncture and the collection of blood samples would be ideally suited to this purpose.

DESCRIPTION OF METHOD

Dr. Silver of IntraCellular Diagnostics, formerly Spectroscan, Inc., devised a system (patented) for collecting epithelial cells from the oral mucosa for the determination of ion concentration. A number of characteristics of these cells influenced the choice for clinical testing. They are non-cornified epithelial cells located on the inferior aspect of the tongue and therefore, well protected from trauma. They have the capability of reflecting relatively recent physiologic changes since they are renewed every 3 days and have aerobic metabolism. Most importantly, they are easily accessible and can be removed by a wooden applicator stick with minimum discomfort. Smears of cells removed in this manner show predominantly individual cells rather than sheets of contiguous cells. This facilitates the visual isolation of single cells with the electron microscope for their analysis. Only intact undamaged cells, approximately 800 cubic microns, are analyzed.

The technique for obtaining the cells and preparing smears is illustrated on Figure la. The mucosa under the tongue is rinsed with distilled water three times. The cells are removed by fairly vigorous scraping of the sublingual area in the groove between the tongue and jawbone, using the pointed end of a wooden curettage stick. The cells are applied to a special slide, and fixed immediately with standard cytology fixative (2.5% carbowax) in 95% ethanol). Smears made in this manner contain cells, dehydrated by the fixative, that can be stored indefinitely at ambient temperatures.

The development of the electron microscope and instruments to control and focus electron beams to irradiate very small areas in cells and subcellular particles gave rise to the technique of x-ray microanalysis that has been applied to the measurement of ions in a number of tissues (2). X-ray emissions that are produced when atoms in a specimen are bombarded by electrons are characteristic of each ion since the charge and energies associated with orbiting electrons are unique for each element. The emission spectra of the elements exposed to preselected energy bands, in the range 500-5000 KeV, are quantified. The intensity of the fluorescence within each energy band is proportional to the intracellular concentration of the ion. Detection of the fluorescence is accomplished with commercially available silicon-lithium detector tubes that can adequately discriminate the energy bands of interest. Six of the nine elements detectable by x-ray microanalysis of sublingual cell are measured in studies of normal physiology: calcium, phosphorus, magnesium, sodium, potassium and chloride. The average concentration of a minimum of 3 cells is the basis for the result in one specimen since the cells in one individual show little variation.

One of the major problems in quantifying the element in a tissue is the background non-specific radiation spectrum produced by x-ray photons generated by incident electron interactions with chemical elements in the analyzed field. To reduce this background, Dr. Silver developed slides or viewing substrates composed of conductive high purity carbon. The fairly constant background emitted by this material is useful in the quantification of the x-ray data that is expressed as a peak to background ratio. Method Summary is illustrated on Figure lb.

APPLICATION OF THE METHOD BY NASA

NASA's principle effort in the development of a test to measure the ion concentration in sublingual cells has been research by the biomedical program carried out by scientists with expertise in skeletal metabolism. These efforts have been directed toward determining the biological meaning and deviations in intracellular ions in non-human primates and in male volunteers for experiments in a model for weightlessness. A brief summary of the experiments and results follows.

Non-human primates were fed synthetic diets containing no vitamin D, the same amount of phosphorus (0.5%) and 2 levels of calcium (1.2 and 0.3%). The Rhesus monkey is known to be highly susceptible to vitamin D deficiency, a problem that might complicate experiments designed to evaluate the response of the skeleton to microgravity. The study in which intracellular ion analysis of sublingual cells was done, was carried out to gather information on the effect of dietary calcium on the biochemical expression of vitamin D deficiency (3). Vitamin D deficient animals showed marked increases in the intracellular concentrations of calcium and phosphorus in sublingual cells that bore no relationship to the changes in serum levels. The result of the intracellular ion assay was consistent with one of the known biologic effects of vitamin D to facilitate the transport of calcium and phosphorus out of cells (4). This experiment provided the biological verification of the test for disorders of calcium and phosphorus metabolism.

The experiments carried out in human adults were designed to establish normative data in healthy 30-55 year old men, the most common candidate for the astronaut corps, and to define the effects of a model of weightlessness, -6 degree head down tilt bed rest, on intracellular ion concentrations. Sublingual cell ion concentrations of calcium, phosphorus and potassium increased by the 3rd bed rest day and remained higher than pre-bed rest levels for the duration of a 28-day study, exceeding the percent changes in the normal variation in concentrations 10 to 50 fold (5). The changes in intracellular ions could be demonstrated before increases in circulating calcium and phosphorus that generally occur on the 6th bed rest day. In one group of subjects who performed a specific type of exercise daily as a countermeasure during a 30-day bed rest study, there were no deviations in the concentration of intracellular ions from pre-bed rest levels (6). Importantly, in these subjects, the common problem of increased urinary calcium excretion and alterations in the hormones of the calcium endocrine system did not take place. We have also obtained cross-sectional data in healthy ambulatory men that relates the levels of sublingual cell calcium, phosphorus, and potassium to the individual's relative level of daily physical activity.

The results from the above experiments provide every indication that the concentration of at least 3 of the ions in sublingual cells accurately reflect an individual's level of musculoskeletal activity. While the mechanism of the observed changes in intracellular ions is not known, the basic explanation may reside in alterations in the calcium endocrine system during musculoskeletal activity that influence the transcellular transport of the calcium, phosphorus and potassium. Future research directed toward these basic questions is needed. For the present, analysis of sublingual cell ion concentrations can be applied to monitoring the adequacy of an individual's exercise program and schedule with confidence. There is enough data to support the design and development of a small instrument for measurements of intracellular ion concentrations in space. With such a system, astronauts could monitor one objective measure of their physical well being in space and use the test results as a guide for planning their daily activity.

Potential applications of sublingual cell analysis for clinical medicine are numerous, but currently limited by the few clinical research studies in patients with well characterized problems. Current research involves analyses in patients with magnesium deficiency and cardiovascular diseases, two areas where estimations of the ion concentrations of cells as an index of mineral and electrolyte metabolism may be more useful in the management of patients than blood tests.

REFERENCES

1. Leach, C. S. and Rambaut, P. C. Biochemical responses of the Skylab Crewmen: an Overview. In "Biomedical Results from Skylab" (R. S. Johnston and L. F. Dietlein, eds), NASA SP-377, pp 204-216.

2. Gupta, B. L., Hall, T. A., and Moreton R. B. Electron Probe X-ray microanalysis. In Transport of Ions and Water in Animals. Gupta, Moreton, Oschman and Wall. Eds. 83-143, Academic Press. London, U.K. 1977.

3. Arnaud, S. B., Jee, W. S. S., Minato, I., Buckendahl, P., Berry, P., Young, D. R. Effect of dietary calcium on the bone of adult vitamin D deficient Rhesus monkeys. J. Bone and Mineral Research 1:S1 #29, 1986.

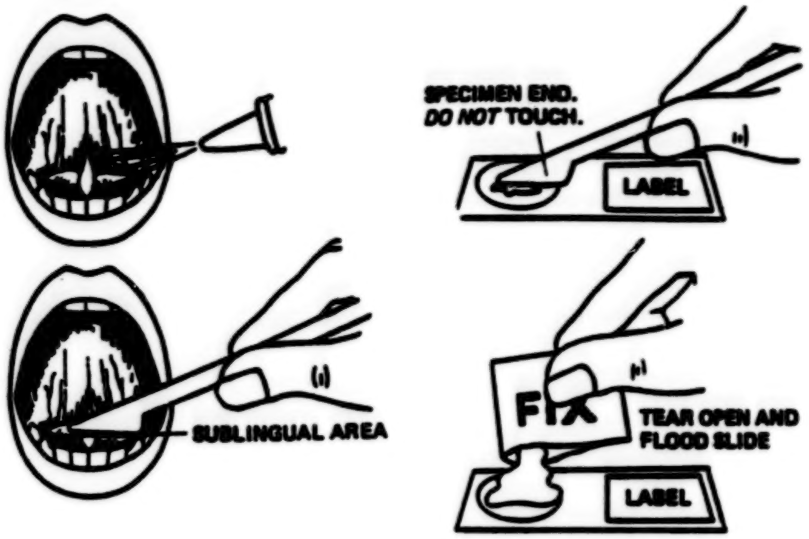
4. Sampson, H. W., Matthews, J. L., Martin, J. H., and Kunin, A. S. An electron microscopic localization of calcium in the small intestine of normal, rachitic, and vitamin D-treated rats. Calcif. Tissue Res. 5:305-316, 1974.

5. Harris, B. A., Silver, B. B., Greenleaf, J. E., Arnaud, S. B. Alterations in intracellular calcium during bedrest with and without exercise. J. Bone and Mineral Research 2: Suppl. 1: 539, #338, 1987.

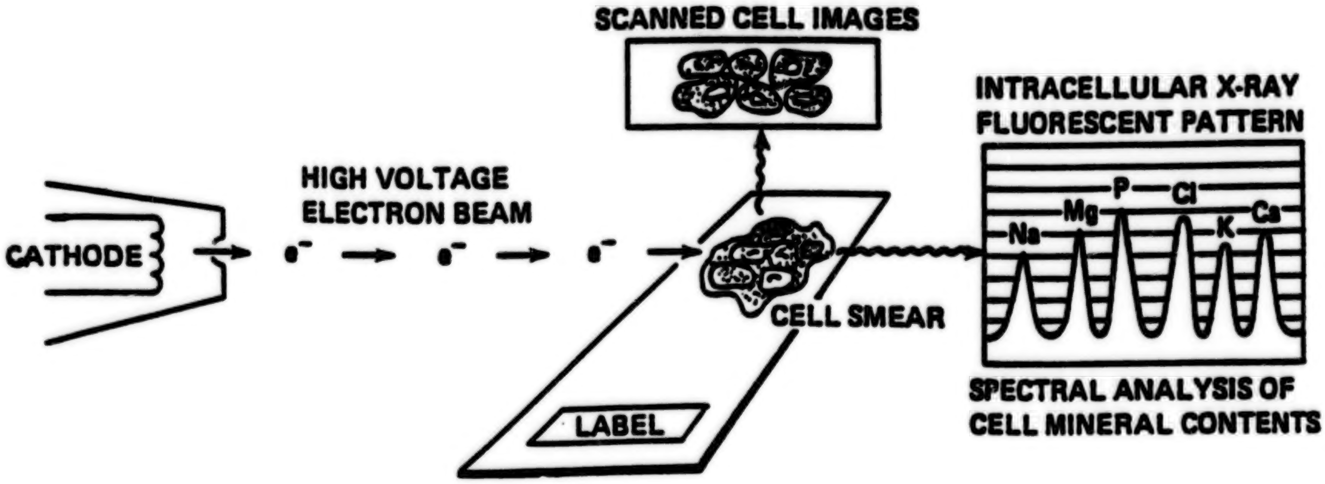
6. Arnaud, S. B., Marcus, R., Greenleaf, J. E. Suppression of the parathyroid/1, 25-dihydroxyvitamin D axis during head down tilt bed rest can be prevented by exercise that loads the skeleton. Endocrine Society, June, 1989. p 470.

FIGURE 1.

a. SAMPLE PREPARATION


SPECIMEN END.
DO NOT TOUCH.
LABEL
(1)
SUBLINGUAL AREA
FIX
TEAR OPEN AND FLOOD SLIDE
LABEL

b. INTRACELLULAR ANALYSIS


SCANNED CELL IMAGES
INTRACELLULAR X-RAY FLUORESCENT PATTERN
HIGH VOLTAGE ELECTRON BEAM
CATHODE
e⁻
e⁻
e⁻
Na
Mg
P
Cl
K
Ca
CELL SMEAR
LABEL
SPECTRAL ANALYSIS OF CELL MINERAL CONTENTS

N91

24059

UNCLAS

N91-24059

OXYGEN PRODUCTION USING SOLID-STATE ZIRCONIA ELECTROLYTE TECHNOLOGY

Jerry W. Suitor
Douglas J. Clark

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

ABSTRACT

High purity oxygen is required for a number of scientific, medical, and industrial applications. Traditionally, these needs have been met by cryogenic distillation or pressure swing adsorption systems designed to separate oxygen from air. Oxygen separation from air via solid-state zirconia electrolyte technology offers an alternative to these methods. The new technology has several advantages over the traditional methods, including reliability, compactness, quiet operation, high purity output, and low power consumption.

INTRODUCTION

The Process

Solid-state zirconia electrolyte technology is made possible by the fact that zirconia is one of a number of materials that is electrically insulating but that can conduct ionic oxygen. Because of this attribute, zirconia can be used as the electrolyte in an electrolytic or galvanic cell. In either case, the zirconia is sandwiched between two gas permeable electronically conducting electrodes. The electrolytic cell, Figure 1, is the basis for the oxygen separation cell.

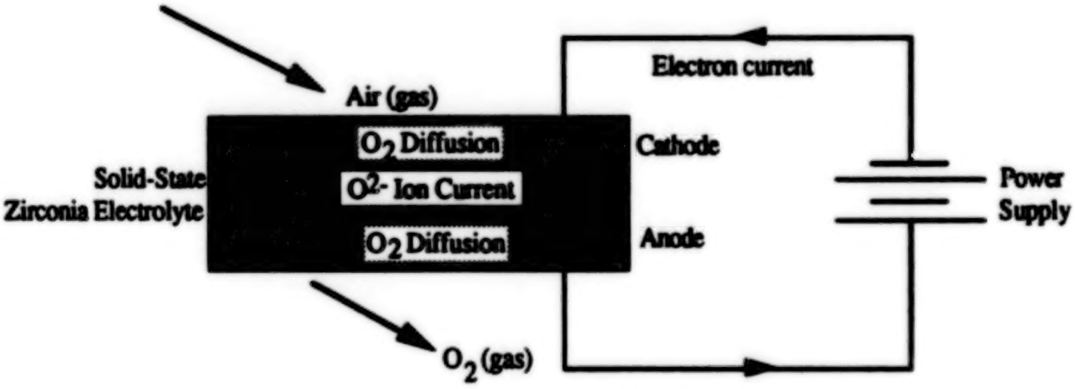


Figure 1. Electrolytic Cell (Oxygen Separation Cell)

The energy necessary to drive the separation process is supplied by the power supply. Oxygen in the air diffuses through the cathode to the cathode-electrolyte interface. Under the influence of the applied voltage, oxygen molecules are dissociated and reduced to oxygen ions, O^{2-}. As an ionic species, oxygen enters the crystal lattice of the electrolyte and moves toward the anode. At the anode, each ion gives up its two electrons which then enter the anode and return to the power supply, completing the circuit. The oxygen atoms recombine into O_2 and diffuse out through the anode. Because the zirconia conducts only oxygen ions, the gas on the output (anode) side of the cell is pure oxygen. Equation 1 describes the reaction.

$$O_2 \text{ (gas 0.2 atm)} + 4e^- \text{ (cathode)} \Rightarrow 2O^{2-} \text{ (electrolyte)}$$

$$2O^{2-} \text{ (electrolyte)} \Rightarrow O_2 \text{(gas 1.0 atm)} + 4e^- \text{ (anode)} \qquad (1)$$

To impart oxygen ion conductivity to the material, pure zirconia (ZrO_2) must be doped with another oxide such as calcia (CaO) or yttria (Y_2O_3). [1] Each dopant cation (Ca^{2+} or Y^{3+}) replaces a Zr^{4+} ion; the net result is that the doped crystal is electrically neutral without having all oxygen ion sites filled. These oxygen ion vacancies impart in the substance an appreciable oxygen ion conductivity. Because the ionic conductivity is a strong function of temperature, the typical operating temperature of solid-state zirconia electrolyte cells is 1000 C.

Thermodynamic Considerations

The laws of thermodynamics play a major role is determining the power consumption of any oxygen separation system.

Oxygen separation in the solid-state zirconia electrolyte oxygen separator is driven by the current provided by the power supply. Part of the overall power consumption is due to the thermodynamically reversible process of separating oxygen from nitrogen and the other gases that make up air. The rest of the power consumption is caused by irreversible processes including electrical resistance in the electrodes, ionic resistance in the electrolyte, and parasitic effects that occur at the electrodes.

The reversible part of the power consumption is manifested by a voltage across the oxygen separation cell called the Nernst voltage. This voltage is a function of the difference in oxygen partial pressure on the two sides of the cell; the greater the difference, the greater the voltage. For a 1000 C cell separating oxygen from air at atmospheric pressure, the Nernst voltage is 44 mV.

Because the ionization reaction takes place inside the cell (at the cathode-electrolyte interface), the structure of the electrodes can have a major effect on cell performance. Gaseous oxygen may diffuse through pores in the electrodes, or it may diffuse through the electrode material itself. Because the presence of electrodes will reduce the rate of oxygen transport to and from the zirconia electrolyte, regions of oxygen starvation and oxygen enrichment are formed at the cathode and anode respectively. The presence of these starved and enriched regions results in an increase in the Nernst voltage. This (irreversible) enhancement of the Nernst voltage is sometimes called gas-phase polarization. A major portion of the development work performed at JPL has been directed toward development of electrode materials that minimize gas phase polarization.

ADVANTAGES OF SOLID-STATE ZIRCONIA ELECTROLYTE OXYGEN SEPARATION

Separation of oxygen from air or other oxygen-containing gases by the solid-state zirconia electrolyte technique has several advantages over traditional methods of separation like cryogenic distillation or pressure-swing adsorption.

Separation of oxygen by solid-state zirconia electrolyte technology lends itself well to modular construction, and thus is easily scaled. JPL has developed patented circular oxygen separation cells that can be stacked, forming a multi-cell stack; multiple stacks can be manifolded together within a common furnace to form a multi-stack module. The modularity inherent in the technology also simplifies maintenance procedures.

Equation 1 indicates that transfer of four electrons is required to conduct each oxygen molecule (O_2) through the electrolyte. Thus it is a simple matter to control the rate of oxygen production by controlling the current applied to the cells. This feature reduces the storage requirements of oxygen separation units and allows operators to better match production to demand. On a small scale, solid electrolyte cells could be used to meter precise amounts of oxygen to partial oxidation processes.

The basic solid electrolyte separation cell has no moving parts. This fact contributes to the high inherent reliability of the technology. The lack of moving parts, combined with the continuous, rather than batch, nature of the process result in a quiet, vibration-free system.

Because the separation of oxygen from air by a solid-state zirconia electrolyte cell is an electrochemical process, the oxygen output stream is 100 percent oxygen. Other separation schemes, such as cryogenic distillation or pressure swing adsorption, cannot produce oxygen of this purity due the basic processes used.

APPLICATIONS OF SOLID-STATE ZIRCONIA ELECTROLYTE TECHNOLOGY

Medical

Solid-state zirconia electrolyte technology has been proposed for several medical applications. Because of the modularity of the technology, application in home, hospital and portable settings are feasible. A block diagram of a home medical unit is shown in Figure 2.

A blower gathers fresh air and forces it through a series of small recuperative heat exchangers. Then, after passing through the startup heater, the air enters the zirconia oxygen separation module. The module separates approximately 50 percent of the oxygen from the input air stream. The oxygen and oxygen depleted air streams pass through heat exchangers before leaving the unit. The heat exchangers thus serve to preheat the input air, reducing the load on the startup heater, and cool the output streams, reducing the amount of waste heat leaving the unit.

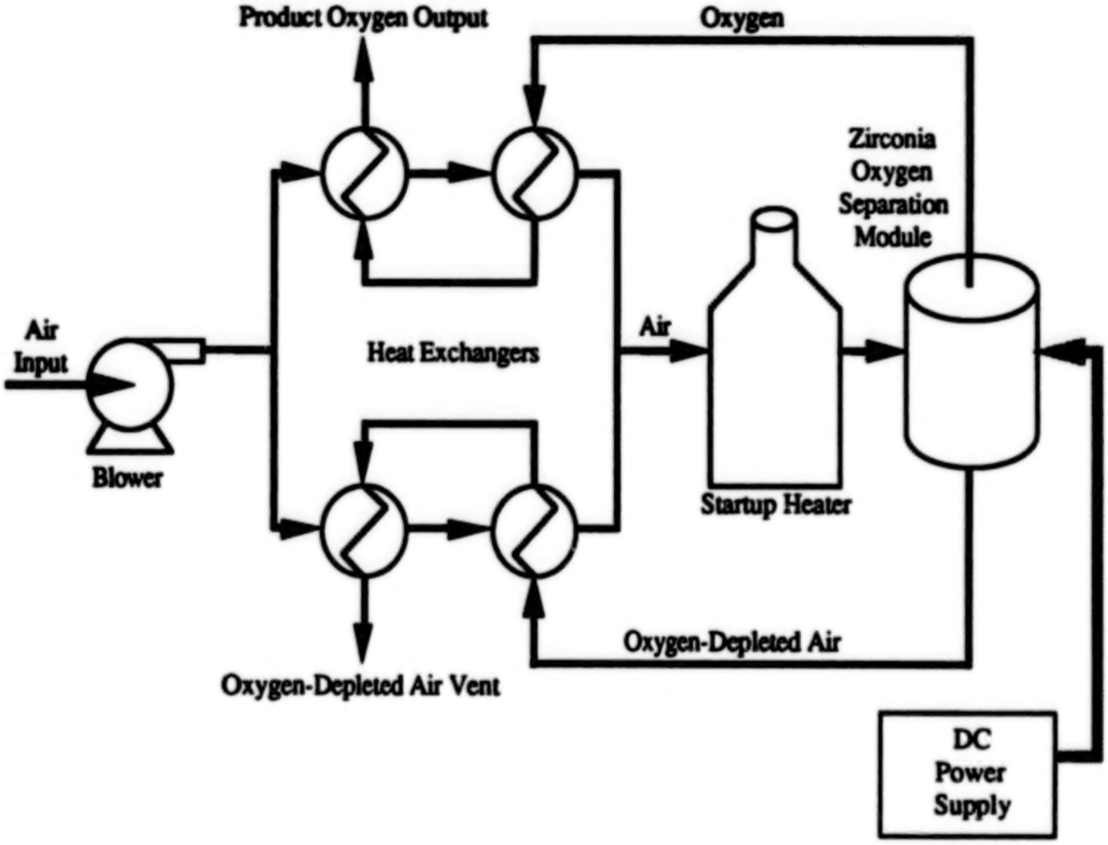

Product Oxygen Output
Oxygen
Zirconia Oxygen Separation Module
Air Input
Heat Exchangers
Air
Blower
Startup Heater
Oxygen-Depleted Air
Oxygen-Depleted Air Vent
DC Power Supply

Figure 2. Home Medical Oxygen Supply Unit

Industrial

Solid-state zirconia electrolyte technology has also been considered in a number of industrial applications. A similar system to that shown in Figure 2 is used in an industrial application. Of course, the size is much larger depending on the quantity of oxygen required. The inherent modularity of the zirconia electrolyte technology permits the construction of units ranging from a few liters/min of oxygen to tons per day. Applications considered for this technology include hospital supply, fish farming, weld shop and metal fabrication, coal gasification, and basic metal refining.

Aerospace

The inherent high reliability of the zirconia oxygen separation process makes the technology attractive for several aerospace applications.

Feasibility studies are currently underway at JPL to determine the suitability of solid-state zirconia electrolyte technology for application in spacecraft-borne sensor coolers. In this application, the zirconia cell would be used as an oxygen pump in a closed-loop Joule-Thomson cooling cycle. The lack of vibration-inducing moving parts makes this scheme especially attractive for sensitive optical systems.

Another application for the technology in space is In-Situ Resource Utilization (ISRU). ISRU refers to the utilization of resources available in space, such as Lunar soil or Martian atmosphere, for propellants, life support, and fabrication. Solid-state zirconia electrolyte technology has been identified as a key element in systems to extract oxygen from Lunar soil air and the Martian atmosphere.

JPL ACCOMPLISHMENTS

Project History

JPL has been working in the field of solid-state zirconia electrolyte technology since 1978. In 1984, the U. S. Department of Energy funded an effort intended to develop oxygen separators for use in coal-fired power plants. JPL made several major accomplishments in the development of this technology during the DOE effort [2, 3]. These accomplishments were focused on cell geometry design, component fabrication techniques, and cell testing.

Cell Geometry

A major factor affecting solid-state zirconia electrolyte oxygen separation cell performance is cell geometry. The presence of ionic resistivity dictates that the zirconia electrolyte should be made a thin as possible. Zirconia tubes are in common use but the wall thickness is very large (~1mm). In this case, electrodes are applied on both the inside and outside surfaces of the tube. Air is passed through the center of the tube, and oxygen is conducted through the tube wall to the outside. In addition to the wall thickness, another major drawback to this geometry is the fact that the supply air is depleted of oxygen as it travels the length of the tube; there is little oxygen left in the air by the time it reaches the far end of the tube. This oxygen starvation increases the oxygen partial pressure difference across the cell and results in an increased Nernst voltage and higher power consumption.

JPL addressed the problem of oxygen starvation by developing a cell geometry based on a circular electrolyte disk. As shown in Figure 3, the supply air enters one side of the cell around the circumference of the disk and is directed toward its center by a number of radial ribs. Oxygen is removed as the air travels toward the center of the disk. Because the cell area available to a given quantity of input air is reduced as the center is neared, the increase in Nernst voltage is not as severe as is seen in tubular designs. Computer models indicating the superior performance of this patented design have been confirmed by experimental data. The ribs support the zirconia permitting very thin (~0.05 mm) membranes to be used. The ribs also serve as busbars to carry current to the cell electrodes. [4]

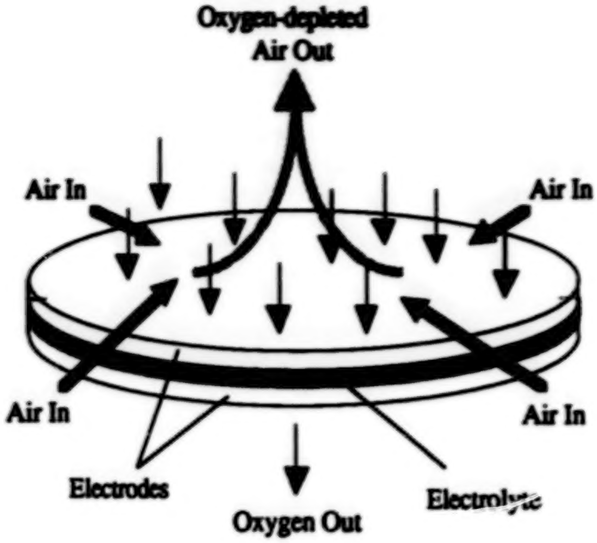


Figure 3. Radial Air Flow, Axial Current Flow Circular Disk Oxygen Separation Cell Geometry (Ribs not shown)

Component Fabrication

JPL developed a patented process for fabrication of the zirconia electrolyte membrane. [5] In this process, conductive ceramic electrodes are applied to the zirconia electrolyte before firing. The method is capable of producing three layer (electrode-electrolyte-electrode) composite oxygen separation membranes with zirconia thickness less than 0.05 mm thick. The firing cycle and temperature were substantially reduced in this patented process. In addition, JPL has developed the ability to fabricate ceramic cell casings.

Testing

As part of the Department of Energy-sponsored effort, JPL built and tested more than sixty single cell oxygen separation units. In addition, ten multicell stacks were tested; three of these were capable of producing 1 standard liter of oxygen per minute each.

Current Work

In addition to promotion of earth-bound applications of the technology, current work at JPL is focussed on studies to determine the suitability of solid-state zirconia electrolyte technology for use in aerospace applications. Two applications under consideration are In-Situ Resource Utilization, in which oxygen would be produced from Lunar soil or the Martian atmosphere, and cryogenic refrigeration, which involves use of a zirconia-based oxygen pump.

TECHNICAL CHALLENGES

The major technological obstacles to commercialization of solid-state zirconia electrolyte technology stem from the high operating temperature required by the process. To eliminate the need for expensive precious metal electrodes and interconnects, JPL's oxygen separation cells use strontium-doped lanthanum manganite (LSM), an electrically conductive ceramic, for both electrodes and cell interconnections. Not only is this material electrically conductive, it is mechanically and chemically stable at 1000 C.

While the problem of expensive interconnects has been eliminated by the use of LSM, sealing of the cells continues to be difficult. The majority of JPL-built separation cells to date have relied upon precision-ground flat seals to contain the the product oxygen. While suitable for use in applications that lack a requirement for pressurized output, the flat seals are incapable of supporting a pressure differential of more than than 4 kPa (0.6 PSI). Follow-on work at JPL will concentrate on development of higher pressure seals.

JPL also intends to examine other oxygen ion-conducting materials such as delta-bismuth oxide. Data reported in the literature indicate that this material has a higher ionic conductivity that doped zirconia, and can operate at a lower temperature.

CONCLUSIONS

JPL has developed an oxygen separation system that is modular and reliable and without moving parts. The power consumption is lower than for other oxygen separation systems. The ceramic fabrication technology has been developed to the point where cells and stacks are easily fabricated. Still required in the development is the successful use of seal technology to provide high pressure, 100 percent purity oxygen to a variety of applications including medical, industrial, and aerospace.

How to License The Technology

Patents held on this technology are the property of the California Institute of Technology. Organizations interested in licensing this technology should contact the California Institute of Technology, Office of Patents and Licensing, (818) 356-4567.

REFERENCES

1. Dell, R. M.; and Hooper, A.: Oxygen Ion Conductors, in Solid Electrolytes: General Principles, Characterization, Materials, Applications, eds. Paul Hagenmuller and W. van Gool. Academic (New York): 1978, 291-312.

2. Suitor, J. W.; Berdahl, C. M.; Ferrall, J. F.; Marner; W. J., Schroeder; J. E.; and Shlichta, P. J.: Development of an Alternate Oxygen Production Source Using a Zirconia Solid Electrolyte Membrane: Technical Progress Report for Fiscal Years 1986 and 1987, Jet Propulsion Laboratory Publication, No. JPL D-4320, May 1987.

3. Suitor, J. W.; Clark, D. J.; Losey, R. W.: Development of Alternative Oxygen Production Source Using a Zirconia Solid Electrolyte Membrane: Technical Progress Report for Fiscal Years 1987, 1988, and 1989, Jet Propulsion Laboratory Publication, No. JPL D-7790, August, 1990.

4. Suitor, J. W.; Berdahl, C. M.; Marner; W. J.: U. S. Patent No. 4,885,142, assigned to Caltech, Patent No. 4,885,142, December 1989.

5. Schroeder, J. E.; and Anderson, H. U.: U. S. Patent No. 4,957,673, assigned to the California Institute of Technology, September 1990.

N91

24060

UNCLAS

N91-24060

MONITORING AND CONTROL TECHNOLOGIES FOR BIOREGENERATIVE LIFE SUPPORT SYSTEMS/CELSS

William M. Knott, Ph-D.
and
John C. Sager, Ph.D.

ABSTRACT

The development of a Controlled Ecological Life Support System (CELSS) will require NASA to develop innovative monitoring and control technologies to operate the different components of the system. Primary effort over the past three to four years has been directed toward the development of technologies to operate a Biomass Production Module. Computer hardware and software required to operate and collect and summarize environmental data for a large plant growth chamber facility has been developed and refined. Sensors and controls required to collect information on such physical parameters as relative humidity, temperature, irradiance, pressure, and gases in the atmosphere; and PH, dissolved oxygen, fluid flow rates, and electrical conductivity in the nutrient solutions are being developed and tested. Technologies required to produce high artificial irradiance for plant growth and those required to collect and transport natural light into a plant growth chamber are also being evaluated. Significant effort has been directed towards the development and testing of a membrane nutrient delivery system potentially useful in growing plants in the microgravity of space. Robotic and plant imaging systems required to manipulate, seed, and harvest crops, and to determine plant health prior to stress impacting plant productivity are also being researched. Tissue culture technologies are being developed for use in management and propagation of crop plants. Though previous efforts have focused on development of technologies required to operate a Biomass Production Module for a CELSS, current efforts are expanding to include technologies required to operate modules such as food preparation, biomass processing, and resource (waste) recovery which are integral parts of a CELSS.

INTRODUCTION

The Controlled Ecological Life Support System (CELSS) is developing the science database required to build and operate an bioregenerative life support system in space. Such a system has been identified to be a requirement to develop a permanent presence in space (Paine, 1986 /1/; Robbins, 1988 /2/). There is insufficient data at this time to predict when such a system will be economically feasible for a specific mission scenario or what components will be included in a functioning CELSS. However, if humans are going to be anything other than visitors in the space environment, such a system must ultimately be developed. The current CELSS program is a research and development effort which should grow steadily into a major hardware development program after a sufficient science and technology database exists.

The CELSS Breadboard Project (Koller, 1986 /3/; Prince and Knott, 1989 /4/) being conducted at the Kennedy Space Center has as its major goal the development and operation of an initial CELSS at a one person scale to demonstrate feasibility for the development of such a system. This project was initiated in late 1986 with the first research data being collected on biomass production in 1988. The current schedule is for the initial Breadboard CELSS configuration to be completed in 1993 and tested by the end of 1994. The components or modules being developed for inclusion in the initial CELSS Breadboard are as depicted in Figure 1. There are four major modules, Biomass Production, Biomass Processing, Food Preparation, and Resource Recovery. For either one of these modules to be operated successfully will require a major enhancement of monitoring and control technologies. This presentation will describe the CELSS Breadboard Project, discuss current technologies being developed, and identify requirements for new technologies necessary for monitoring and controlling the system.

THE BREADBOARD PROJECT

The current Breadboard Project is centered around a large atmospherically sealed plant growth facility, the Biomass Production Chamber (BPC) (Figure 2), which was constructed to measure energy use and mass flow through a crop community over its entire life cycle. A series of experiments are being conducted in the BPC to evaluate crop species in this closed environmentally controlled facility (Wheeler and Sager, 1990 /5/; Wheeler et. al, 1990 /6/). During these trials, emphasis is being placed on the effect of different environmental conditions on plant growth, CO_2 exchange, water use, and mineral uptake. Microbiological sampling and analyzes are conducted during each plant growth study in order to determine the functional relationships between the microorganisms and the plants, especially in the rhizosphere. Total biomass production is measured for each crop as the plants are harvested from the chamber. Other modules of a CELSS which are required to convert this biomass into food and to recycle all materials back to the plants are being developed in laboratories adjacent to the BPC. The edible biomass is transferred to the processing laboratory (kitchen) so that meals can be prepared and tested. Various processing procedures are being evaluated in order to determine amount of time required to accomplish the task, to identify equipment requirements, and to evaluate the meals for nutritional value and palatability. The inedible biomass is transferred to the resource recovery lab so that various subcomponents that may convert this material into an edible substance can be evaluated. These subcomponents include aquaculture, enzyme degradation of cellulose into sugars, and fungal production processes. Material left over from either the food preparation or biomass processing activities must be converted to CO_2 , water, and minerals for return to the Biomass Production Chamber. Resource recovery subcomponents that are currently being tested to accomplish this include a leachate reactor and aerobic microbiological reactors.

Tests of each resource recovery subcomponent is designed to obtain mass flow data through each component in the context of a total integrated system. Emphasis is being placed on the cycling of carbon, hydrogen, oxygen and nitrogen along with selected major minerals. A related activity will evaluate total pyrolysis of the edible biomass to determine the products generated and the energy used. The measurement of water cycled through each subcomponent of the Breadboard is of major interest during all project activities. The water loop on the BPC is currently being closed and hygiene water being used by the human along with his liquid waste will be measured and recycled through the appropriate subcomponents.

The initial total Breadboard system is scheduled to be complete by 1993. Extended testing of this system will occur during 1993 and 1994. A total reconstruction of a CELSS, the Integration Test Facility, at a three to four person scale is scheduled for 1995.

ADVANCED TECHNOLOGY DEVELOPMENT

One of the major problems encountered to date in the construction and operation of the Breadboard CELSS is the reliability of the physical systems required to maintain the biological organisms. The biology has been very reliable and predictable but the monitoring and controlled technologies required to maintain a proper environment for the organisms has been unreliable. Several objectives need to be met as we attempt to improve the technologies required for the successful deployment and operation of a CELSS in space. Monitoring and control technologies must be developed to minimize energy use and limit manpower requirements. The elements must be miniature in size, light in mass, and durable. Their measuring capabilities must be real time, on-line, and adequately sensitive for proper system controls. Their operational reliability must be over an extremely long time duration and must require minimal maintenance. Finally, elements of a CELSS which will be deployed in free space must have the capability of operating in a microgravity environment.

Several advanced technologies required to operate the Biomass Production Module of a CELSS have been identified. Fiber optic sensors for on-line monitoring and control for selected parameters in the BPC are being developed and tested. Parameters to be measured by the sensors include atmospheric ethylene, trace organic contaminants, pH, moisture levels, microbial organisms, and specific ions in nutrient delivery solutions. These sensors use primary and secondary absorbance and liquid atomic emission technologies as their primary measuring processes. Most of the sensors currently available for monitoring these and other parameters in the Biomass Production Chamber are not sensitive enough or when on-line require excessive maintenance to keep them operative. Development of computer software will be required in order to establish expert control of the atmosphere in the BPC using the outputs from the array of sensors. We have developed some chamber control software along with new data display capability for the BPC during the past two years. Technologies for the production, transportation, and distribution of irradiance for plant growth is another primary area of interest to the CELSS program. Current research in this area has concentrated on a light pipe material and fiber optics for transportation and distribution of the light energy. Of major importance in this technology area is the increased efficiency of bulbs and the maximum delivery of photons to the plant surface.

At least one project is evaluating noninvasive remote sensing systems that may be useful in detecting water stress and/or nutrient deficiency in plants. A technology that will detect plant stress prior to it reaching a level to cause a decrease in productivity is vital to the successful development of a CELSS. Robots with vision imaging capabilities will be required to locate sensors and conduct plant manipulations such as seeding and harvesting. A seeding robot for wheat has been developed and tested. The development of robotic capabilities that will significantly reduce the time required to operate the Plant Production Module is important. Finally, if the CELSS is to operate in a microgravity environment then a major requirement is to develop a system to deliver water and nutrients to the roots of plants under these conditions (Wright, et. al. 1988 /7/). We have developed and tested a membrane nutrient delivery system(Figure 3) as a potential technology to accomplish this task (Dreschel and Sager, 1989 /8/). Other projected areas requiring technology development in order to develop a successful Biomass Production Module for a CELSS include genetic plant engineering, management of tissue culture systems, trace contaminant removal from air and water, and water collection devices.

Many of the technologies discussed previously for the Biomass Production Module are also required for the Food Preparation, Biomass Processing, and Resource Recovery Modules. Advance sensor technology and expert computer control are needed to operate bioreactors that are vital to these modules. The requirements for these sensors and computer software are identical to or only slightly modified from those discussed for the Biomass Production Module. New equipment to thrash, mill, and grind edible biomass and to process the same into food needs to be developed. This equipment must be small, easy to operate, and extremely efficient so that little, if any, wastage occurs. The aquaculture system will require some advance technologies that have not been previously discussed. Sensors and controls for the nitrogen wastes produced by the fish must be developed. Development of a food for the fish using only inedible plant material is a major challenge. Technologies to reduce the amount of water to support the fish and video systems to monitor the health of the fish are a few other capabilities which require research. There is no doubt as these processing and resource recovery modules become more a part of the Breadboard Project, additional advance technologies required to operate these modules will be identified.

SUMMARY

The CELSS Breadboard Project is proceeding on schedule and a demonstration of feasibility at a one person scale should be accomplished in the next 3-4 years. The Biomass Production Module is complete and producing data on plant growth and development, water cycling, and carbon dioxide and oxygen exchange. During the past 2-3 months, efforts have been initiated to develop the Food Preparation, Biomass Processing, and Resource Recovery Modules that will complete the initial CELSS Breadboard. In order for a CELSS to be deployable in space to meet the life support requirements of humans, significant advancement in selected technologies is required. A few technologies are currently being researched such as fiber optic sensors, noninvasive remote plant stress sensors, expert computer controls, and microgravity compatible nutrient delivery systems. Improved sensor sensitivity and reliability, efficient irradiance production and delivery systems, and expert computer controls must be developed in order to control the physical systems required to maintain the biology. Technologies to monitor biological stress, to accomplish genetic engineering, and to provide tissue culture maintenance must be completed prior to the development of an operational CELSS. The development of advanced technologies are primarily aimed at reduction in energy use, limiting manpower requirements, miniaturization of subcomponents, improved operation reliability, on-line monitoring capabilities, and functionality in microgravity. In order for human beings to develop a truly permanent presence in space, a CELSS must be developed and a significant advancement in technologies is required in order to make such a life support system possible.

LITERATURE

1. Pioneering the Space Frontier. National Commission on Space (The Paine Commission), New York, Bantam Books, 1986.

2. Exploring the Universe: A Strategy for Space Life Sciences. (The "Robbins" Report), Life Sciences Strategic Planning Study Committee, NASA Advisory Council, Washington, D.C.

3. Koller, A. M.. CELSS Breadboard Facility Project Plan. Biomedical Operations and Research Office. National Aeronautics & Space Administration, Kennedy Space Center. March 1986.

4. Prince, R. P. and Knott, W. M.: CELSS Breadboard Project at the Kennedy Space Center. In: D. W. Ming and D. L. Henninger (eds.), Lunar Base Agriculture; Soils for Plant Growth. Amer. Soc. Agron., Madison, WI USA. 1989.

5. Wheeler, R. M. and Sager, J. C.: Carbon Dioxide and water exchange rates by a wheat crop in NASA's Biomass Production Chamber results from an 86-day study (January to April 1989). NASA Technical Memorandum. TM 102788. January 1990.

6. Wheeler, R. M., Mackowaik, C. L., Dreschel, T. W., Sager, J. C., Prince, R. P., Knott, W. M., Hinkle, C. R. and Strayer, R. F.: System development and early biological tests in NASA's Biomass Production Chamber. National Aeronautics & Space Administration. Technical Memorandum. TM 103494. March 1990.

7. Wright, B. D., Bausch, W. C. and Knott, W. M.: A Hydroponic system for microgravity plant experiments. Trans of the ASAE. 31(2). 440-446. 1988.

8. Dreschel, T. W. and Sager, J. C.: Control of water and nutrients using porous tube: A method for growing plants in space. HortScience 24(6): 944-947. 1989.

CELSS BREADBOARD CONCEPT

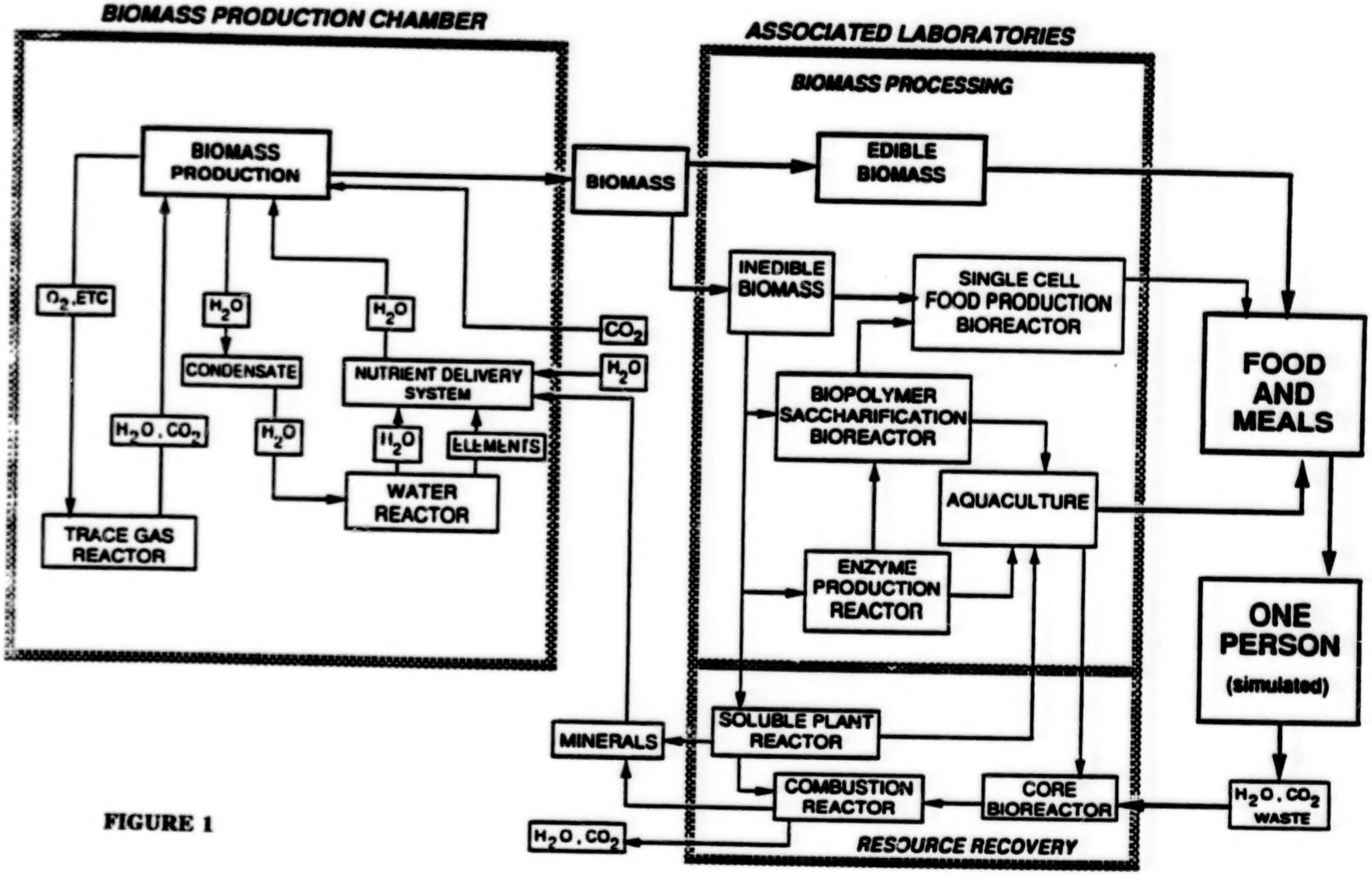
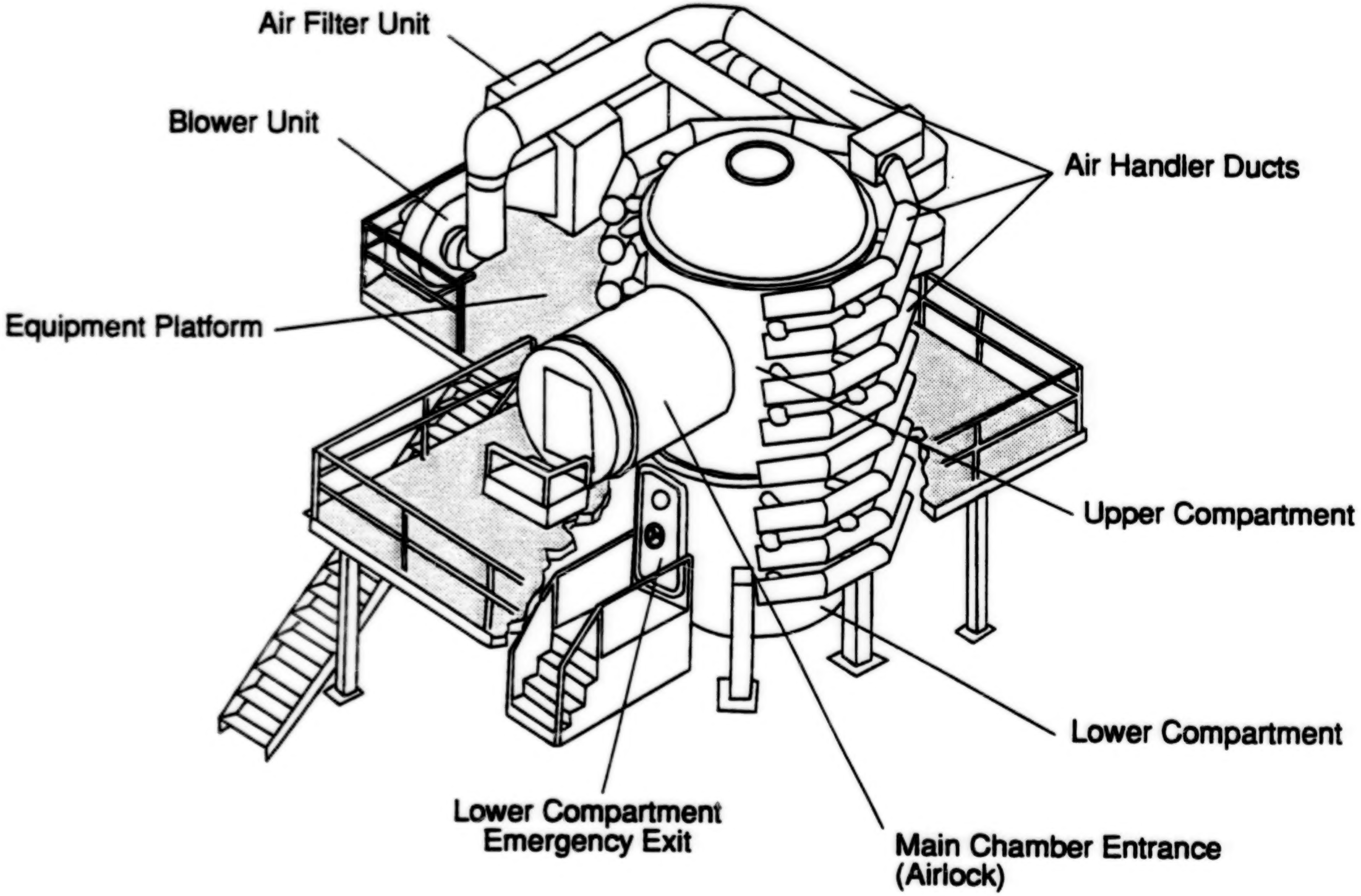

BIOMASS PRODUCTION CHAMBER
BIOMASS PRODUCTION
O_2,ETC
H_2O
H_2O
CONDENSATE
NUTRIENT DELIVERY SYSTEM
H_2O, CO_2
H_2O
H_2O
ELEMENTS
TRACE GAS REACTOR
WATER REACTOR
BIOMASS
CO_2
H_2O
MINERALS
H_2O, CO_2
ASSOCIATED LABORATORIES
BIOMASS PROCESSING
EDIBLE BIOMASS
INEDIBLE BIOMASS
SINGLE CELL FOOD PRODUCTION BIOREACTOR
BIOPOLYMER SACCHARIFICATION BIOREACTOR
AQUACULTURE
ENZYME PRODUCTION REACTOR
SOLUBLE PLANT REACTOR
COMBUSTION REACTOR
CORE BIOREACTOR
RESOURCE RECOVERY
FOOD AND MEALS
ONE PERSON (simulated)
H_2O, CO_2 WASTE

FIGURE 1


Air Filter Unit
Blower Unit
Equipment Platform
Air Handler Ducts
Upper Compartment
Lower Compartment
Lower Compartment Emergency Exit
Main Chamber Entrance (Airlock)

The Biomass Production Chamber

FIGURE 2

POROUS TUBE PLANT GROWTH UNIT

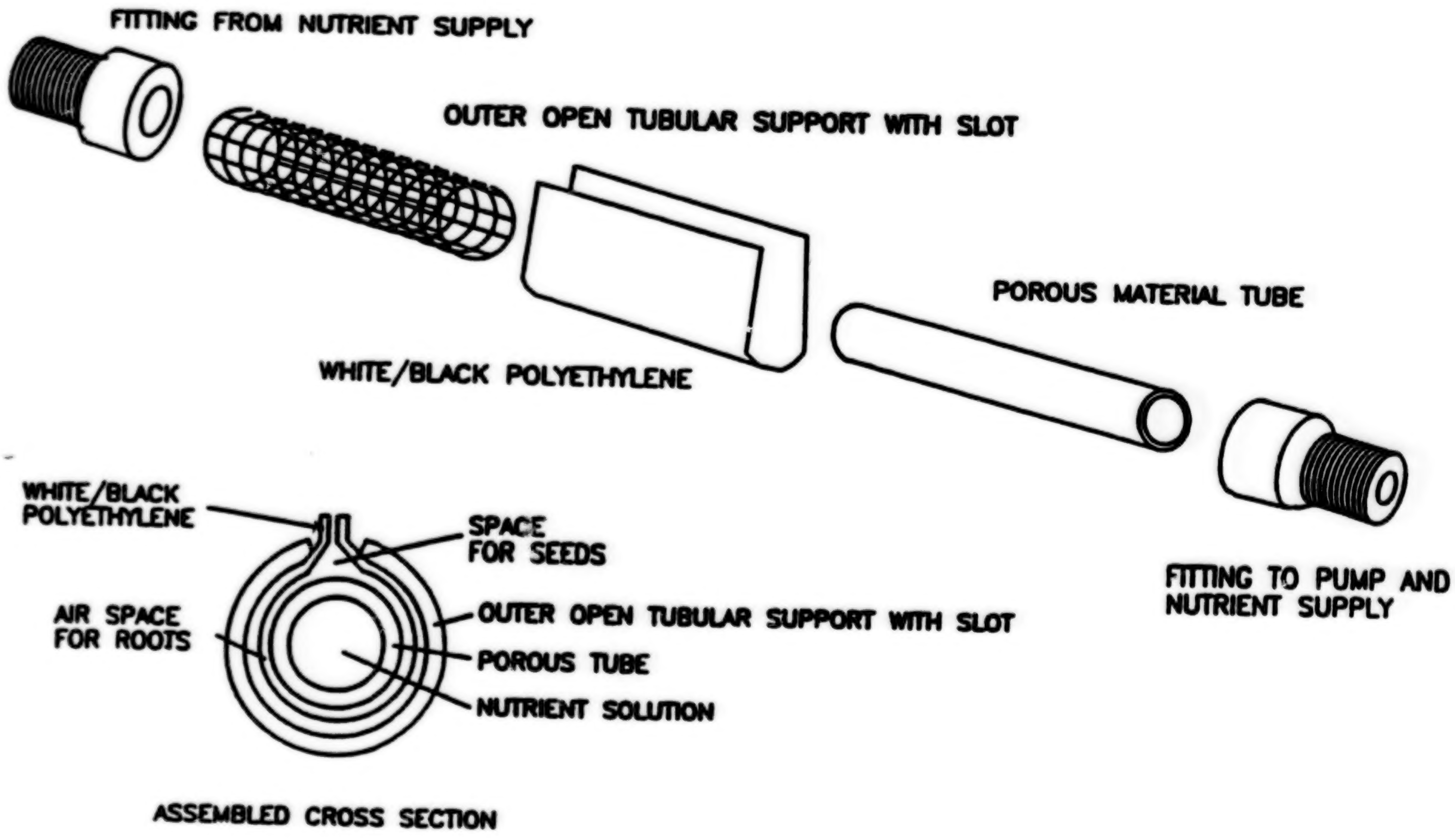

FITTING FROM NUTRIENT SUPPLY
OUTER OPEN TUBULAR SUPPORT WITH SLOT
POROUS MATERIAL TUBE
WHITE/BLACK POLYETHYLENE
FITTING TO PUMP AND NUTRIENT SUPPLY
WHITE/BLACK POLYETHYLENE
SPACE FOR SEEDS
AIR SPACE FOR ROOTS
OUTER OPEN TUBULAR SUPPORT WITH SLOT
POROUS TUBE
NUTRIENT SOLUTION
ASSEMBLED CROSS SECTION

FIGURE 3. SCHEMATIC DIAGRAM OF THE SECOND DESIGN OF THE POROUS TUBE PLANT GROWTH UNIT

SESSION E - MATERIALS SCIENCE (PART 2)

Wednesday November 28, 1990

- Silicon Carbide, An Emerging High-Temperature Semiconductor
- Flexible Fluoropolymer-Filled Protective Coatings
- A Conformal Oxidation-Resistant, Plasma-Polymerized Coating
- Flame-Retardant Composite Materials
- Superplastic Forming Of Al-Li Alloys For Lightweight, Low-Cost Structures
- Localized Corrosion Of High-Performance Metal Alloys In An Acid/Salt Environment

PRECEDING PAGE BLANK NOT FILMED

N91
24061
UNCLAS

N91-24061

SILICON CARBIDE, AN EMERGING HIGH TEMPERATURE SEMICONDUCTOR

Lawrence G. Matus, Ph.D.
Research Scientist

J. Anthony Powell
Senior Research Physicist

NASA Lewis Research Center
Cleveland, OH 44135

ABSTRACT

In recent years, the aerospace propulsion and space power communities have expressed a growing need for electronic devices that are capable of sustained high temperature operation. Applications for high temperature electronic devices include development instrumentation within engines, engine control and condition monitoring systems, and power conditioning and control systems for space platforms and satellites. Other earth-based applications include deep-well drilling instrumentation, nuclear reactor instrumentation and control, and automotive sensors.

To meet the needs of these applications, the High Temperature Electronics Program at the Lewis Research Center is developing silicon carbide (SiC) as a high temperature semiconductor material. Research is focussed on developing the crystal growth, characterization and device fabrication technologies necessary to produce a family of silicon carbide electronic devices and integrated sensors. This paper will present the progress made in developing silicon carbide and discuss the challenges that lie ahead.

INTRODUCTION

In recent years, the aerospace propulsion and space power communities have acknowledged the growing need for electronic devices that are capable of sustained high temperature operation. Applications for high temperature electronic devices include development instrumentation within engines such as multiplexers analog-to-digital converters, and telemetry systems capable of withstanding hot section engine temperatures in excess of 600° C. Similarly, engine mounted integrated sensors and electronics could reach temperatures which exceed 500° C while uncooled operation of control and condition monitoring equipment in advanced sustained supersonic aircraft would subject electronic devices to temperatures in excess of 300° C (1). Hypersonic vehicles will ultimately pose even more severe demands on electronic devices and sensors.

In addition to aeronautics, there are many other areas that would benefit from the existence of high temperature electronic devices. Space applications include power electronic devices for Space Station Freedom, space platforms, and satellites. Since power electronics require radiators to dissipate waste energy (heat), electronic devices that are capable of operating at higher temperatures would allow a reduction in radiator size. This results in a weight savings and thereby reduces the cost of placing the hardware into orbit.

The need for electronic devices capable of sustained operation at high temperature is not restricted to the aerospace community. Earth-based applications include deep-well drilling instrumentation, power electronics for motor control, nuclear reactor instrumentation and control and automotive electronics and sensors.

To meet the needs of the applications mentioned above, the High Temperature Electronics Program at the Lewis Research Center is developing silicon carbide (SiC) as a high temperature semiconductor material. This program supports two major elements of the Center's mission: to perform basic and developmental research aimed at improving (1) aerospace propulsion, and (2) power systems. Research is focussed on developing the crystal growth, crystal characterization and device fabrication technologies necessary to produce a family of SiC devices.

SILICON CARBIDE: THE SEMICONDUCTOR

Silicon Carbide (SiC) is familiar to most as the abrasive grit material on sandpaper. It is, however, a material that possesses many other useful properties. Crystalline SiC can be found in refractory, structural and electrical applications as well as the common abrasive applications. Due to its extreme hardness (only diamond and boron carbide are harder) SiC grit is used extensively in lapping, grinding, cutting, and polishing operations. As a refractory material, SiC is used as a protective coating on components exposed to high temperature and/or corrosive environments. In the aerospace industry, SiC fibers are of interest as a high strength, low density reinforcement material.

It is the semiconducting properties of electronic grade SiC crystals that make it particularly attractive for high temperature applications. This is because SiC possesses a wide energy bandgap. The maximum operating temperature for a semiconductor is determined by the forbidden bandgap energy. The useful temperature limit is reached when the number of intrinsic carriers, thermally excited across the energy gap, approaches the number of purposely added (extrinsic) carriers. This temperature (when expressed as the absolute temperature) is roughly proportional to the energy bandgap. Depending on the particular polytype (structural form) of SiC, the bandgap energy varies from 2.2 eV to 3.3 eV.

Based on the inherent solid state properties of silicon (bandgap energy of 1.1 eV), the maximum temperature at which a silicon device could theoretically operate is 300^0 C. Conventional silicon electronic devices are rated to operate at temperatures up to 125^0 C, i.e., the MIL-SPEC (military specifications) limit. Since the desired operating temperature for some of the applications mentioned above approaches 600^0 C, it is clear that a new semiconductor material will have to be developed. Another choice as a high temperature semiconductor based upon its bandgap energy might be gallium arsenide which could theoretically operate at 460^0 C, but it has not proven operationally stable at this temperature. Using the same criteria as applied to silicon, SiC could theoretically be employed as a semiconductor at temperatures as high as 1200^0 C. A more reasonable, shorter term goal is to produce electronic devices capable of 600^0 C operation.

In comparing potential candidate materials for high temperature semiconductor devices, SiC stands out not only because of its excellent high temperature electronic properties but also because it is a very stable ceramic material up to temperature of 1800^0 C. The combination of the material's high thermal conductivity (heat transfer) and high breakdown field (tolerant of high electric fields) provides the potential for improved power electronic systems and for increasing the number of devices per unit area. Those properties which determine the high frequency characteristics of semiconductor devices also appear to be excellent for SiC and superior to those of silicon or gallium arsenide.

Component reliability is a key issue in all aerospace applications because failure can lead to expensive or tragic consequences. Electronic devices or sensors that are capable of operating at high temperatures have the immediate payoff of improved reliability when operated at lower temperatures. For example, if electronic devices capable of 300^0 C operation possess the same failure rate at 300^0 C as devices specified for 125^0 C operation, the failure rate will be reduced by 1000 when the "300^0C" electronic devices are operated in a 125^0 C environment. This three orders of magnitude improvement in reliability is due to the exponential dependence of failure rates on temperature. Based on its properties, the reliability of electronic devices and sensors fabricated from SiC should be much higher than that obtainable from any current semiconductor material.

SILICON CARBIDE CRYSTAL GROWTH

With all the advantages that SiC possesses, why has SiC technology not been incorporated into electronic systems? The main problem has been the lack of SiC single crystals suitable for device fabrication purposes. Until recently, there was no process whereby single crystals of SiC with sufficient size, purity, and perfection could be grown reproducibly. SiC does not melt at any reasonable temperature and pressure conditions so this rules out the "growth-from-melt" technique commonly used to obtain other semiconductor single crystals such as silicon and gallium arsenide.

Historically, vapor phase growth processes have proven to be the most successful method for producing SiC crystals. Early research was done on SiC crystals that were a by-product of the industrial Acheson process for making sandpaper grit and abrasives (2). In the Acheson process, SiC is formed at 2400° by the reaction of silica and coke. At this temperature, gas pockets can form within the SiC reaction product. The SiC sublimes and then condenses on the inside walls of gas pockets located at cooler parts of the reaction product. Occasionally, isolated SiC crystals are produced within these pockets during the production process. The larger and better crystals were hand selected for electrical research purposes.

In 1955, Lely developed a laboratory version of the industrial sublimation process and was able to produce rather pure SiC crystals (3). Encouraged by the Lely process, NASA Lewis and other laboratories pursued the development of SiC semiconductor devices during the 1960's and early 1970's. Though SiC devices were demonstrated above 400° C, by the early seventies the Lely process and other processes had not matured to the point where high-quality large-area crystals could be grown reproducibly. Since crystal substrates are crucial to device fabrication, interest in SiC waned, and from 1973 to 1980, there was very little effort in the U.S. on SiC. However, research did continue in Japan and in Europe during this period.

In 1980, because of the increased need for high temperature electronics in advanced turbine engines, NASA Lewis again embarked on a high temperature electronics program. The emphasis again has been on developing SiC. The problem regarding the crystal growth of SiC is rooted in the fact that SiC crystals can take on many different structural forms called polytypes. The many polytypes of SiC (over 140) differ from one another in the stacking sequence of the SiC double layers. Depending on the stacking, either cubic, hexagonal or rhombohedral structures are possible. The two most common SiC polytypes are 3C-SiC and 6H-SiC representing the cubic and hexagonal structures, respectively. In the early research, SiC crystals grown by sublimation techniques were a mixture of different polytypes. Since each SiC polytype has its own electronic properties (i.e., bandgap energy, carrier mobility, etc.), sublimation grown SiC crystals usually contain heterojunctions and possess unpredictable and non-uniform electronic properties.

To favor growth of a single polytype of SiC, epitaxial growth on a host crystalline substrate from gases containing silicon and carbon was hypothesized. The host crystal imparts its crystalline regularity to the thin growing layer. Since silicon is available in perfect, large, and low-cost crystals, many attempts at the heteroepitaxial growth of SiC on Si were made. These efforts were largely unsuccessful because of the large lattice mismatch that exists between Si and SiC (e.g., the SiC lattice is 20% smaller that the Si lattice).

Large area heteroepitaxial growth of 3c-SiC on Si was finally achieved at the NASA Lewis Research Center in 1982 by using a chemical vapor deposition (CVD) process (4). Crystal growth takes place at atmospheric pressure in a fairly conventional horizontal CVD system. A complete system description is given in reference 5. The CVD reaction system is illustrated schematically in Figure 1. To grow a single crystal 3C-SiC, first an electronic grade silicon substrate is placed on an rf-heated graphite susceptor. The essential step in the growth process is a rapid temperature ramp from near room temperature to a growth temperature of 1360° C in the presence of a hydrocarbon gas. The NASA Lewis Research Center process uses propane as the carbon containing gas. During the first two minutes of growth, a single crystal 3C-SiC film about 20-nm-thick is produced on the Si substrate. After this initial SiC growth, silane gas is added to provide a silicon source for the final step, the bulk growth of 3C-SiC to the desired thickness. During this time, the 3C-SiC layer grows at a rate of 3 to 4 μ m/hr.

3C-SiC is a transparent, yellow crystal which fractures into regular rectangular pieces. Although visually the material appears to be of relatively good quality, in actuality a high density of defects exists in the crystal. Certain defects can adversely affect the electrical properties of SiC devices. During the past eight years, much progress has been made in understanding problems associated with 3C-SiC grown on Si but much research remains to be done in this area.

A recent development in SiC crystal growth is having an enormous impact on SiC research. A SiC research team at North Carolina State University announced the successful implementation of a seeded-growth sublimation method to produce the 6H-SiC polytype in boule (large cylinder) form (6). A private company, Cree Research, Inc., has developed this process to the point where 1-inch diameter wafers of 6H-SiC crystals can now be used as substrates for SiC epitaxial growth via the CVD processes already developed. Growth of high quality 6H-SiC epitaxial films has now been achieved at NASA Lewis using 6H-SiC wafers as substrates (7). Figure 2 is a photograph comparing a Cree Research Inc. SiC wafer with available Lely SiC crystals. Prior growth experiments had been performed on irregular-shaped Lely or Acheson SiC crystals but the small size and uncertain quality of these crystals make them unsuitable for commercial production.

Doping (intentional insertion of the extrinsic carriers) the epitaxial films with electrical impurities to produce n-type and p-type SiC is vital to the realization of electronic devices. Addition of nitrogen gas to the growth process gases results in nitrogen incorporation into the SiC lattice. Since nitrogen is a donor impurity in SiC, n-type SiC is produced. To produce p-type SiC, aluminum has been used as an acceptor impurity. Aluminum is incorporated by adding trimethylaluminum to the growth process gases.

CHARACTERIZATIONOF SiC FILMS

Initially, the transition from the Si substrate to the 3C-SiC epitaxial layer was thought to occur by means of a thin buffer layer or transition layer of the order of 20-nm-thick (4). However, high resolution transmission electron microscopy (TEM) has demonstrated that the SiC/Si interface is abrupt with no transition region (8). the 3C-SiC films do contain a large density of defects that include interfacial twins, stacking faults, and inversion domain disorder (9). The defect density in the films is greatest near the SiC/Si interface and decreases with distance away from the interface.

A particular type of lattice defect, called inversion domain boundaries (IDB's) was reported to be present in all SiC films grown on Si (10). The IDB's can be made visible by chemical etching, sputter etching, wet oxidation or 3C-SiC growth in the presence of diborane (11). IDB's form in the initial stages of growth on the Si substrate when SiC islands of opposite phase nucleate and grow together. Across the IDB, the chemical bonding is between like atoms (i.e., Si-Si or C-C), instead of the normal Si-C bond between neighboring atoms.

Normally, in epitaxial growth on Si, the surface of the Si substrate is oriented precisely parallel to an atomic plane, e.g., the (001) plane. It had been found in the growth of gallium arsenide on Si, that IDB's are eliminated by orienting the substrate slightly off-axis from the (001) plane. This technique was applied to SiC growth at NASA Lewis with the result that for Si substrates that were tilted 1° to 4° from the (001) plane, all IDB's were eliminated from the films grown. In addition, the resultant SiC films were smoother by a factor of 2 to 3 than the films grown on-axis substrates (11). Further work is needed to determine the effect on electrical properties and the performance of devices fabricated from these films. Also, the optimum tilt angle and the direction of tilt have yet to be determined.

Electrical characterization to determine electrical properties of the SiC films is an important and necessary evaluation step if high quality SiC films are to be achieved. To determine the carrier concentration and carrier mobility, room temperature Hall measurements were made on n-type 3C-SiC films grown at NASA Lewis using the van der Pauw configuration (12). Ohmic electrical contacts consisted of sputtered tantalum followed by sputtered gold. In order to perform a detailed analysis of the charge carrier concentration, three 3C-SiC films were selected for Hall measurements over the temperature range 50-300K. Experimental results for the films studied showed that the films are highly compensated with a ratio of acceptor (p-type) to donor (n-type) atoms of 0.90. At this time, the identity of the donor and acceptor impurities is not clear. Crystal defects such as vacancies, interstitial atoms, antisite atoms, stacking faults, or dislocation may be acting as donor and acceptor impurities.

It is believed that all present day 3C-SiC films grown at NASA Lewis and elsewhere are compensated. The consequence of compensation is degradation of device quality. Compensated semiconductors have many ionized impurities (positive for donors and negative for acceptors) embedded in the crystal lattice and these ions serve as scattering centers for moving charge carriers. This increases the impurity scattering and reduces the total mobility of the charge carrier compared to uncompensated material with the same density of free carriers. Identification and eventual elimination of the compensating impurities remains a research goal.

As mentioned above, the 3C-SiC films epitaxially grown on Si substrates contain many lattice defects indicating that Si is not the perfect host substrate, at least not when using the current growth processes. The availability of 6H-SiC substrates for epitaxial growth of SiC films is now steering the growth research away from 3C-SiC growth on Si and toward 6H-SiC growth on 6H-SiC.

The defect density in 6H-SiC films grown on 6H-SiC substrates is reduced by a minimum of three orders of magnitude compared to the 3C-SiC films. This is expected to have a very positive effect on SiC device characteristics. As more 6H-SiC films are produced, characterization techniques that have been developed for 3C-SiC films will be systematically employed. At this time, structural characterization (cross sectional transmission electron microscopy), optical characterization (low-temperature photoluminescence), and electrical characterization of both n and p type 6H-SiC films are underway.

Sic DEVICE FABRICATION TECHNOLOGY

The availability of SiC substrates now allows NASA Lewis to place increasing emphasis on device fabrication. In-house research is pursuing the fabrication of in-situ grown junction diodes and metal-insulator-semiconductor field effect transistors (MISFETs).

Junction diodes were produced by first growing an 8-micrometer-thick n-type 6H-SiC film, and then growing an additional 0.75-micrometer-thick p-type 6H-SiC film with trimethylaluminum added to the process gases. Aluminum was incorporated into the growing SiC film to produce p-type material. An array of diode mesa structures was then formed by photolithography followed by reactive ion etching using sulfur hexafluoride and oxygen gases. Electrical contacts, aluminum to the p-type and gold/tantalum to the n-type, were applied by sputter deposition.

A typical current-voltage (I-V) curve for one of the 6H-SiC grown-junction diodes at room temperature and at 600° C is shown in Figure 3. the function of a diode is to allow current to pass in one direction (the forward direction), but not in the opposite (reverse) direction. Hence, an ideal I-V curve would be nearly vertical in the forward direction, and nearly horizontal in the reverse direction. As seen in the I-V curves, the electrical characteristics of the SiC diode change very little when operated from room temperature to 600° C. This is a significant result because the diode (e.g., p-n junction) is a basic building block for electronic devices. In addition to rectification, junctions are also used extensively as isolation layers. For example, an n-channel field-effect-transistor (FET) can be fabricated by first growing a p-type layer followed by an n-type layer. Current flow will then be confined to the n-layer if the voltage polarities are chosen so as to reverse bias the junction.

A depletion-mode metal-oxide-semiconductor field-effect-transistor (MOSFET) was fabricated using the principles of channel isolation as described above. In an FET structure, the current flow is controlled by applying a voltage to the gate electrode. For an n-channel FET, a negative voltage applied to the gate will deplete the channel of electrons and thus "pinch-off" the current flow. In this manner, an FET resembles a switch. The switch is turned on and off by the application of the gate voltage. Figure 4 shows the I-V characteristics of the MOSFET at room temperature and at 500° C. Although the electrical characteristics of the FET are not ideal, the achievement of transistor I-V characteristics at 500° C is an extremely important starting point.

CONCLUDING REMARKS

Ultimately, the goal of the NASA Lewis High Temperature Electronics Program is to develop SiC integrated circuits and monolithic sensors with compensating and signal conditioning electronics integrated into the sensor structure. Electronic devices and sensors that are capable of operating at elevated temperatures eliminate or reduce the amount of cooling that is required. In many space applications, this can be a significant weight savings. Other payoffs of high temperature integrated sensors include reduced cabling and shielding requirements and development of distributed control architectures, i.e., smart actuators.

The development of semiconductor materials does not occur over night. First germanium, then silicon and then gallium arsenide, for example, have come to the marketplace after years in the laboratory and many dollars spent for development cost. The history of SiC as a high-temperature semiconductor has been one of high expectations followed by disappointment. Recent advances in crystal growth of SiC and the increased knowledge of the bulk material properties of the grown SiC are cause for renewed enthusiasm. Although the development of SiC falls into the category of high-risk research, the future looks very promising and the potential payoffs are tremendous. SiC now appears ready to emerge as a useful semiconductor material.

REFERENCES

1. Nieberding, W. C. and Powell, J. A.: IEEE Transactions on Industrial Electronics, Vol. IE-29, May 1982, p.103.
2. Acheson, A. G.: British Patent No. 17911 (1892).
3. Lely, J. A.: Ber. Dt. Kerm. Gas, Vol. 32, 1955, p. 229.
4. Nishino, S., Powell, J. A., and Will, H. A.: Appl. Phys. Lett., Vol. 42, No. 5, March 1, 1983, p. 460.
5. Powell, J. A., Matus, L. G., and Kuczmarksi, M. A.: J. Electrochem. Soc., Vol. 134, No. 6, June 1987, p. 1558.
6. Carter, Jr., C. H., Tang, L., and Davis, R. F.: Presented at the Fourth National Review Meeting on the Growth and Characterization of SiC, Raleigh, NC, 1987.
7. Powell, J. A., Larkin, D. J., Matus, L. G., Choyke, W. J., Bradshaw, J. L., Henderson, L., Yoganathan, M., Yang, J., and Pirouz, P.: Appl. Phys. Lett., Vol. 59, No. 15, April 9, 1990, p. 1442.
8. Chorey, C. M., et al: Semiconductor-Based Heterostructures: Interfacial Structure and Stability, M. L. Green, et al, eds., The Metallurgical Society, Warrendale, PA, 1986, p. 115.
9. Pirouz, P., et al: Heteroepitaxy on Silicon II, J. C. C. Fan, et al, eds., MRS Symp. Proc. Vol. 91, Materials Research Society, Pittsburgh, PA, 1987, p. 399.
10. Pirouz, P., Chorey C. M., and Powell, J. A.: Appl. Phys. Lett., Vol. 50, No. 4, January 26, 1987, p. 221.
11. Powell, J. A., Matus, L. G., Kuczmarksi, M. A., Chorey, C. M., Cheng, T. T., and Pirouz, P.: Appl. Phys. Lett., Vol. 51, No. 11, September 14, 1987, p. 283.
12. Segall, B., Alterovitz, S. A., Haugland, E. J., and Matus, L. G.: Appl. Phys. Lett., Vol. 49, No. 10, September 8, 1986, p. 584

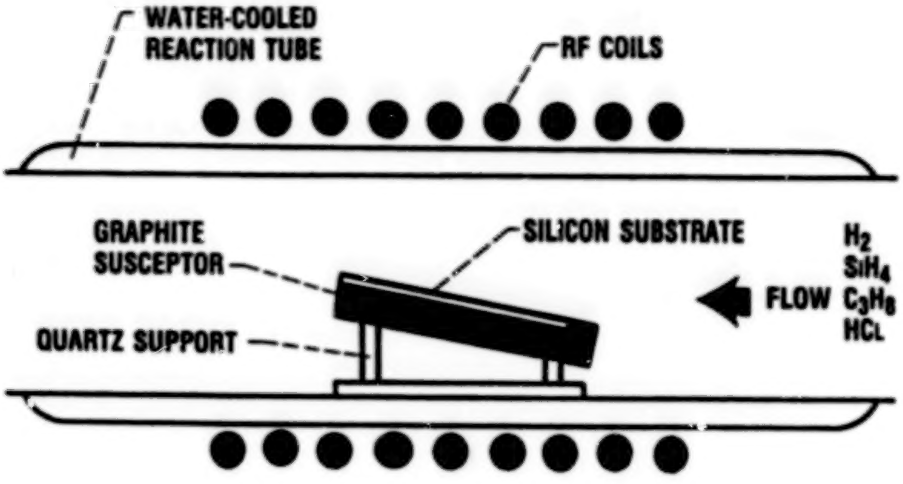

WATER-COOLED
REACTION TUBE
RF COILS
GRAPHITE
SUSCEPTOR
SILICON SUBSTRATE
QUARTZ SUPPORT
FLOW
H_2
SiH_4
C_3H_8
HCl

FIGURE 1. SCHEMATIC DIAGRAM OF REACTION CHAMBER FOR EPITAXIAL SiC CRYSTAL GROWTH

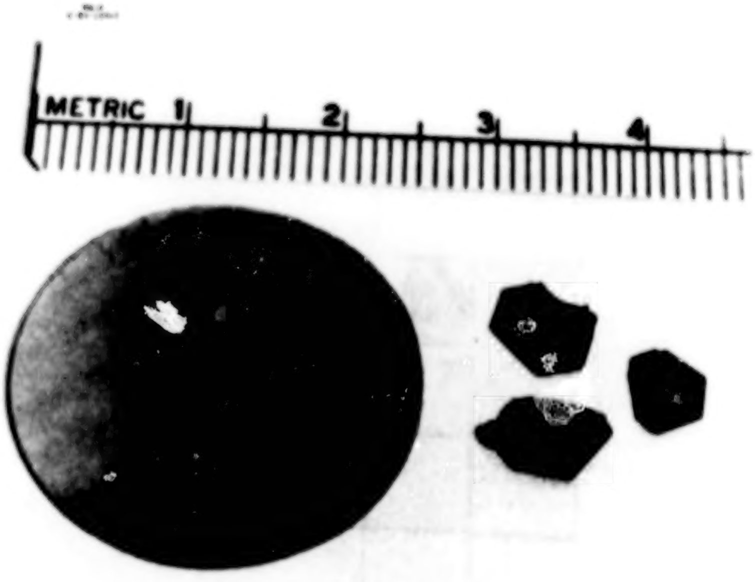

METRIC
1
2
3
4

FIGURE 2. LEFT: CREE RESEARCH INC. 6H-SiC WAFER
RIGHT: LELY 6H-SiC CRYSTALS

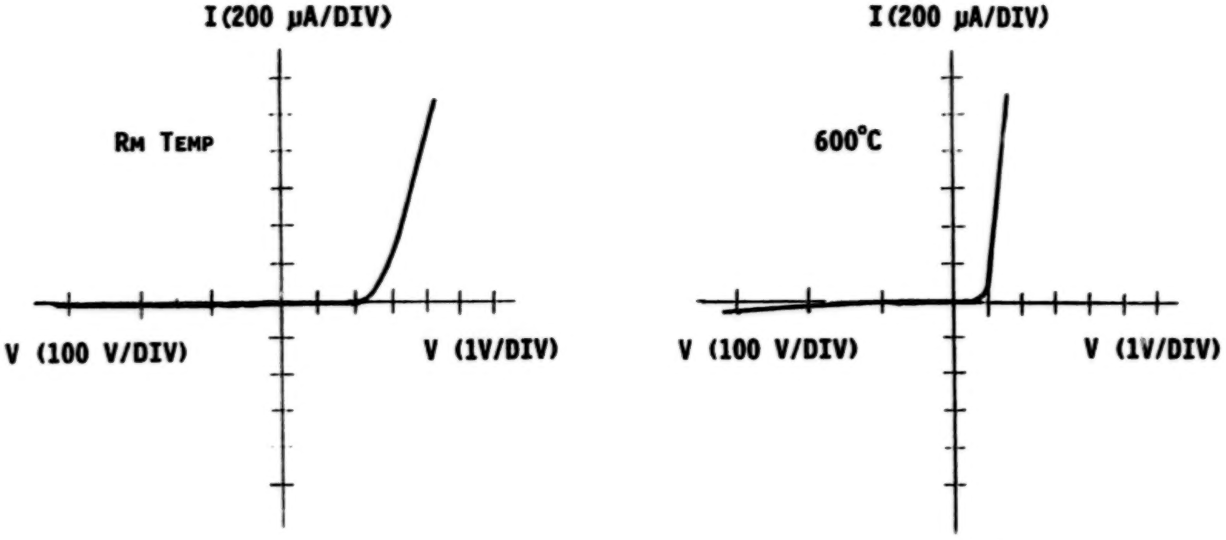


FIG. 3A

FIG. 3B

FIGURE 3. CURRENT-VOLTAGE (I-V) CURVES FOR A 6H-SiC GROWN JUNCTION DIODE

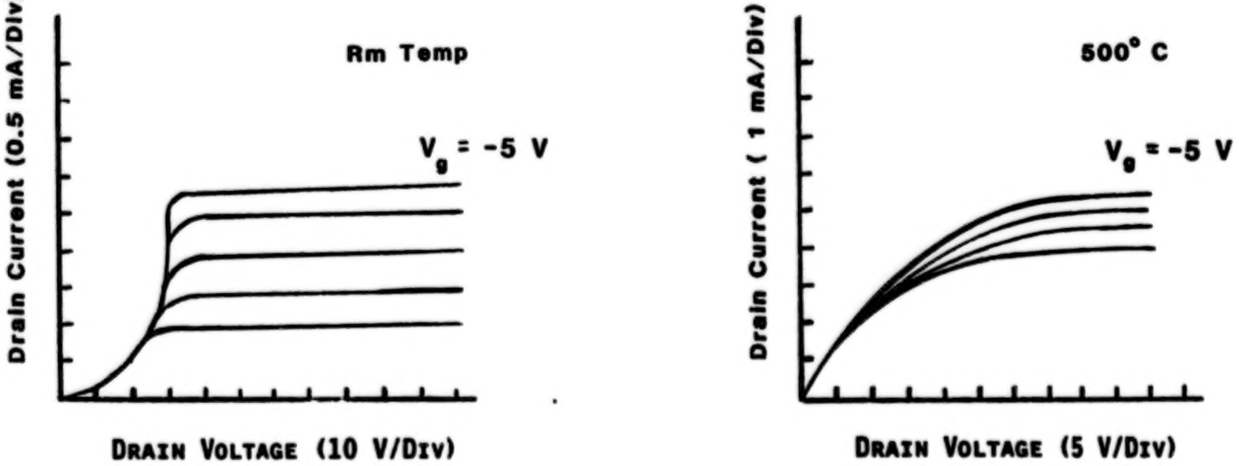


FIGURE 4. CURRENT-VOLTAGE (I-V) CURVES FOR A 6H-SiC DEPLETION-MODE MOSFET

N91

24062

UNCLAS

N91-24062

FLEXIBLE FLUOROPOLYMER FILLED PROTECTIVE COATINGS

Bruce A. Banks, Michael J. Mirtich, James S. Sovey,
Henry Nahra, and Sharon K. Rutledge

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

ABSTRACT

Metal oxide films such as SiO_2 are known to provide an effective barrier to the transport of moisture as well as gaseous species through polymeric films. Such thin film coatings have a tendency to crack upon flexure of the polymeric substrate. Sputter co-deposition of SiO^2 with 4%-15% fluoropolymers has been demonstrated to produce thin films with glass-like barrier properties that have significant increases in strain-to-failure over pure glass films, thus improving their tolerance to flexure on polymeric substrates. Deposition techniques capable of producing these films on polymeric substrates are suitable for durable food packaging and oxidation/corrosion protection applications.

Introduction

Polymeric materials used on spacecraft in the low-earth-orbital environment have been shown to degrade through oxidation because of reaction with the environmental atomic oxygen (ref. 1). Atomic oxygen in the low-earth-orbital environment is formed by photodissociation of the diatomic oxygen by solar photons whose wave length is less than 2430Å. Atomic oxygen is a predominant species between the altitudes of 180 and 650 km (ref. 2). Polymers have not yet been developed which are durable to atomic oxygen interaction. As a result, current approaches to prevent atomic oxygen degradation have utilized protective coatings over the oxidizable substrates. These coatings have consisted of metal oxides such as SiO_2 and Al_2O_3, as well as mixed metal oxide/fluoropolymer films (refs. 1,3 and 4). Because metal oxides are typically in their highest oxidation state, they act as ideal atomic oxygen protective barriers to prevent oxidation of the underlying polymeric materials. However, pure metal oxide films such as SiO_2 have a very limited strain-to-failure, thus decreasing their utility on substrates which are subjected to flexure. The addition of a small fluoropolymer content to the predominantly SiO_2 films greatly increases the strain-to-failure of the films while still maintaining their protective properties. Numerous aerospace spinoff applications exist such as: hermetic coatings, protective transparent coatings, and non-stick coatings. this paper will present deposition techniques for the application of flexible thin film fluoropolymer filled SiO_2 protective coatings, as well as discuss film properties and potential applications.

Deposition Technique

Molecularly mixed metal oxide/fluoropolymer films can be produced by simultaneous sputter etching of sputter targets consisting of silicon dioxide and polytetrafluoroethylene (PTFE-Teflon) (Ref. 3). Figure 1 shows the ion beam sputter deposition configuration for co-deposition of SiO_2 or Al_2O_3 with a fluoropolymer. this apparatus consists of an electron bombardment ion source operating on 500-1000 eV argon ions to sputter-etch targets of metal oxide simultaneous with the polytetrafluoroethylene (PTFE-Teflon). As a result of the simultaneous bombardment of both types of materials, a molecular mixture of metal oxide and fluoropolymer scission fragment species deposits on surfaces exposed to the sputter ejecta from these targets. Because the ion beam current density distribution is typically normally distributed, the ratio of deposition species can be readily controlled by utilizing a pie-shaped polytetrafluoroethylene target segment on top of a circular SiO_2 sputter target (ref. 5). Figure 2 shows a typical plot of the volume fraction of fluoropolymer in the deposited SiO_2/fluoropolymer-mixed film as a function of the polytetrafluoroethylene target angle. The deposition can be performed by either ion beam or RF magnetron deposition with the substrate at room temperature. The deposition rates produced by ion beam sputter-deposited films are typically lower than those which may be produced by RF magnetron deposition processes.

Thin Film Properties

Although silicon dioxide and aluminum oxide are known to be durable to atomic oxygen attack, pure fluoropolymer deposited films are oxidizable by atomic oxygen attack. However, small quantities of fluoropolymers can be molecularly mixed with metal oxide films and still provide atomic oxygen protection if the thin film thickness is slightly increased over that of the pure metal oxide film. Figure 3 shows the minimum thickness fluoropolymer-filled SiO_2 film necessary for permanent atomic oxygen protection as a function of fluoropolymer content fraction in the SiO_2 film (ref. 3). Thinner protective coatings do not provide durable protection thus allowing atomic oxygen reaction with the underlying polymer. For atomic oxygen protective coatings in space, the film thicknesses typically between 650-1500Å are used.

Fluoropolymer filled SiO_2 protective coatings appear colorless and quite transparent on polymeric substrates. Figure 4 illustrates the differences in reflectance, absorptance, and transmittance of SiO_2 films containing up to 4% fluoropolymer sputter deposited on 127μm thick polyimide Kapton.

The value of adding the small addition of fluoropolymer to the metal oxide films is to increase the strain to which the silicon dioxide can be subjected without brittle fractures developing. Figure 5 illustrates the increase of strain-to-failure by small additions of fluoropolymer to the film as well as the minimum radius curvature which film can survive when deposited on a 127μm thick polyimide Kapton substrate without brittle fracture. As can be seen from Figure 5, a fluoropolymer fill fraction of 15% will increase the strain-to-failure by a factor of three. In addition, thin polymeric materials with a 15% fluoropolymer fill fraction can be bent around a radius of curvature of approximately 1mm without brittle fracture. However as can be seen from Figure 3, such films must be at least 600Å to provide atomic oxygen oxidation protection. Oxygen diffusion through thin films appears to be limited to not more than 300Å based on isotope labeled diffusion studies conducted in reference 6. However, it appears that 15% fluoropolymer-filled SiO_2 films whose thicknesses exceed 600Å are quite flexible and atomic oxygen durable. Sputter deposition of these co-deposited films has been shown to produce an intrinsic stress which can tend to deform the substrate polymers. Figure 7 is a plot of the intrinsic stress of co-deposited thin films on silicon substrates for various composition films and film thicknesses (ref. 7).

Atomic oxygen durability of pure SiO_2 and 4% fluoropolymer-filled SiO_2 films were evaluated in space and found to reduce oxidation by a factor of 3,7500 over that of unprotected polyimide Kapton (ref. 7).

Potential Application

One of the primary space applications considered for such films is for the protection of flexible polyimide Kapton solar array blankets that would be used in low earth orbit. the fluoropolymer content of such films may enable inflatable structures to be protected from oxidation as well. Polymeric materials which must be subjected to flexure or small radius of curvature would be ideally suitable for use with fluoropolymer filled films because of their tolerance to high strain in comparison to that of pure metal oxide films.

Because the fluoropolymer-filled SiO_2 films are clear, flexible, corrosion-resistant, chemically inactive, and an effective moisture barrier, there may be numerous non-aerospace applications. One of the most obvious applications is the packaging of food products where it is desirable to prevent diffusion of water vapor through the thin film polymeric container. Generally acceptable long-term hermetic protection is achieved by metallization of plastic film packages. However, such containers do not allow visual inspection of the content of the containers. The flexible fluoropolymer-filled SiO_2 films would be both visibly transparent and offer a hermetic protection in excess of that of uncoated polymeric packaging. The fluoropolymer content of the film may be useful in cases where a slight reduction in coefficient of friction is also important. There may be obvious non-stick applications where the durability of the SiO_2 component can be utilized in concert with the non-stick fluoropolymer component for such applications as cookware.

The chemical properties of the fluoropolymer ion beam sputter-deposited component of the films is thought to be similar to that of polytetrafluoroethylene based on the deposition of pure fluoropolymer films from polytetrafluoroethylene targets (ref. 8). The sputter-deposited fluoropolymer films are very similar to that of the bulk polytetrafluoroethylene except the sputter-deposited fluoropolymer is slightly more cross-lined, which produces a higher thermal decomposition temperature (ref. 8). Applications where abrasion-resistance, oxidation protectiveness, and transparency are requirements include durable acrylic coatings for surface protection of floor coverings, sky light transparency protective coatings, and protection of polycarbonate face shields. These oxidation protective coatings, their deposition process, and associated apparatus have been patented by NASA under U.S. Patent Numbers 4,664,980,4, 560,577and 4,604,181 respectively.

SUMMARY

Molecularly mixed fluoropolymer-silicon dioxide films cannot be fabricated by conventional bulk melting or mixing processes. However, these mixed films can be deposited by ion sputter co-deposition, as well as magnetron sputter deposition processes. Fluoropolymer contents up to 15% by volume enable a factor of three (3) gain in strain-to-failure of the silicon dioxide films, thus allowing a much greater flexibility in the films to be applied on thin polymeric substrates. The co-deposited films are clear, atomic oxygen durable, and pose potential for hermetic coatings for applications ranging from food stuff packaging to face shields. The fluoropolymer content in these coatings may be advantageous to non-stick applications as well.

REFERENCES

1. Banks, B. A., et al: Sputtered Coatings for Protection of Spacecraft Polymers. NASA TM-83706, 1984.

2. Banks, B. A., et al: Atomic Oxygen Undercutting of Defects on SiO_2 Protected Polyimide Solar Array Blankets. Presented at the Materials Degradation in Low Earth Orbit Symposium of the 119th TMS Annual Meeting and Exhibit, Anaheim, CA, 1990.

3. Banks, B. A., et al: Protection of Solar Array Blankets from Attack by Low Earth Orbital Atomic Oxygen. Presented at the Eighteenth IEEE Photovoltaic Specialists Conference, Las Vegas, NV, 1985.

4. Banks, B. A., et al: Performance and Durability of Space Solar Dynamic Power System Optical Surfaces. Presented at the ASME International Solar Energy Conference, Miami, FL, 1990.

5. Nahra, H.: Protective Coatings for Polymers Exposed to Atomic Oxygen in the Low Earth Orbits. M.S. Thesis, Cleveland State University, Cleveland, OH, 1983.

6. Gulino, D. A., et al: Isotopic Study of Oxygen Diffusion in Silicon Dioxide Thin Films. Published in the Journal of Thin Solid Films, 188 (1990) 237-246.

7. Banks, B. A., et al: Ion Beam Sputter-Deposited Thin Film Coatings for Protection of Spacecraft Polymers in Low Earth Orbit. NASA TM-87051, 1985.

8. Banks, B. A., et al: Ion Beam Sputter Etching and Deposition of Fluoropolymers. NASA TM-78888, 1978.

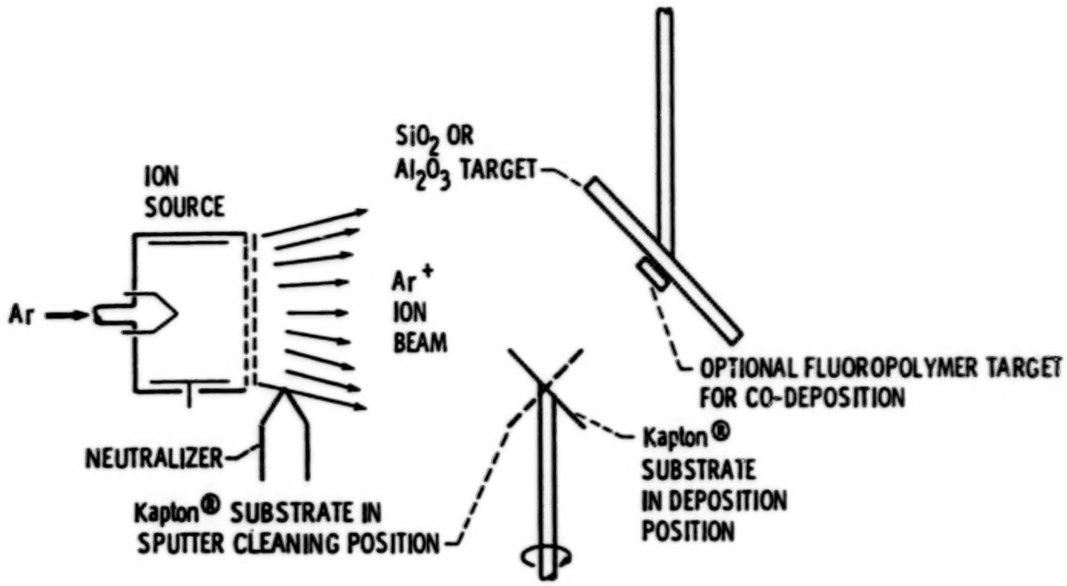

SiO$_2$ OR
Al$_2$O$_3$ TARGET
ION
SOURCE
Ar
Ar$^+$
ION
BEAM
OPTIONAL FLUOROPOLYMER TARGET
FOR CO-DEPOSITION
Kapton®
SUBSTRATE
IN DEPOSITION
POSITION
NEUTRALIZER
Kapton® SUBSTRATE IN
SPUTTER CLEANING POSITION

Figure 1. Ion beam sputter deposition system for molecularly mixed metal oxide fluoropolymer thin films.

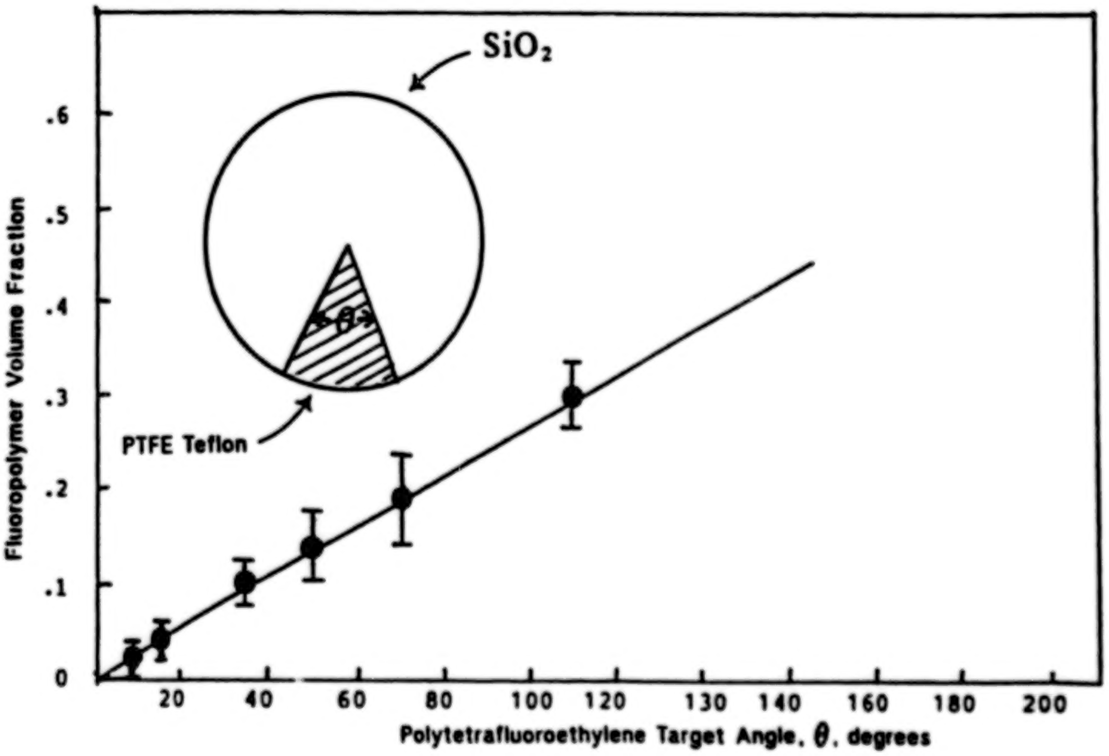


Figure 2. Fluoropolymer volume fraction of fluoropolymer filled SiO$_2$ sputter deposited films.

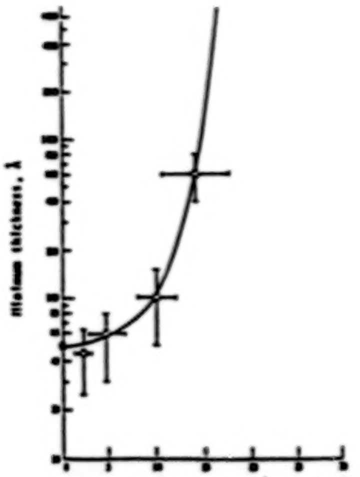


Figure 3. Minimum film thickness for atomic oxygen protection as a function of fluoropolymer content in fluoropolymer film SiO_2 sputter deposited films.

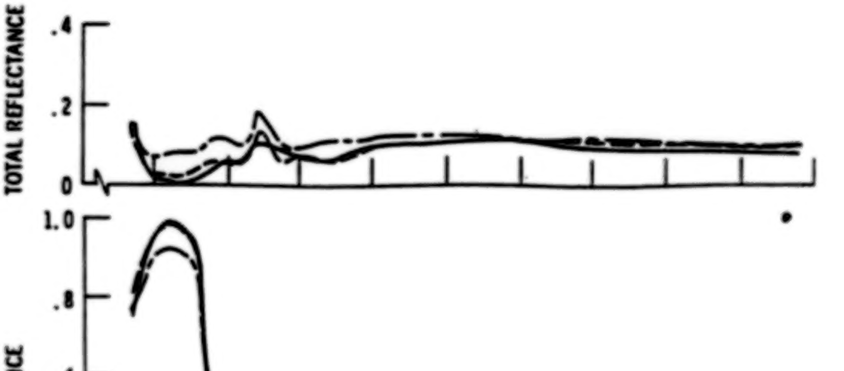
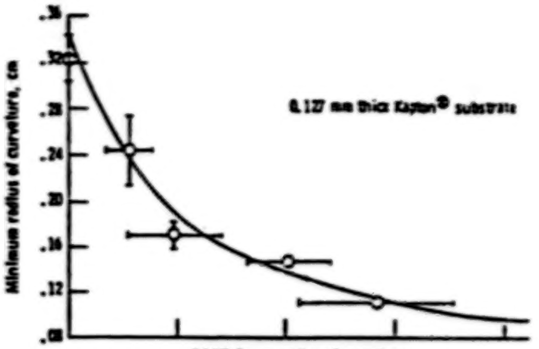
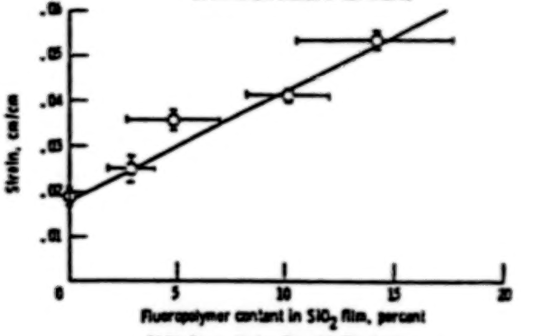


Figure 4. Optical properties of ≥ 96% SiO_2 ≤ 4% PTFE coated Kapton samples unexposed and exposed to low-earth-orbital environment compared with uncoated and unexposed Kapton.



(a) Minimum radius of curvature.



(b) Maximum strain without brittle fracture.

Figure 5. Minimum radius of curvature and maximum strain that a co-deposited SiO_2-fluoropolymer film (1000Å) can survive without brittle failure.

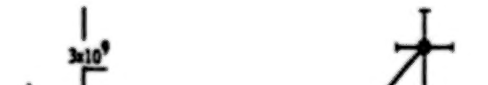


Figure 6. Intrinsic stress of co-deposited thin films on silicon substrates.

N91

24063

UNCLAS

N91-24063

A CONFORMAL OXIDATION-RESISTANT, PLASMA-POLYMERIZED COATING

Morton A. Golub, Theodore Wydeven, and Narcinda R. Lerner
NASA Ames Research Center
Moffett Field, California 94035

ABSTRACT

A comparative study was made of the surface recession (etching) of thin films of plasma-polymerized tetrafluoroethylene (PPTFE), polytetrafluoroethylene (PTFE), and ion-beam sputter deposited polytetrafluoroethylene (SPTFE) exposed to ground-state atomic oxygen [O(^{3}P)] downstream from a nonequilibrium radio-frequency O_2 plasma. At 22°C, the etch rates for PTFE, SPTFE, and PPTFE were in the ratio of 8.7:1.8:1.0. A thin, conformal coating of PPTFE (etch rate of 0.3 nm/h at 22°C) was found to protect an underlying cast film of a reactive polymer, *cis*-1,4-polybutadiene, against O(^{3}P) attack for the time required to fully etch away the PPTFE coating. From ESCA analysis, PTFE exhibited only minor surface oxidation (uptake of 0.5 atom % O) upon etching, its F/C ratio decreasing slightly from 2.00 to 1.97; PPTFE exhibited considerable surface oxidation (uptake of 5.9 atom % O) and a decrease in F/C ratio from 1.30 to 1.23; and SPTFE exhibited a surface oxidation (uptake of 2.2 atom % O) intermediate betweeen those of PTFE and PPTFE, with a decrease in F/C ratio from 1.73 to 1.67. A plasma-polymerized fluorocarbon coating such as PPTFE might be useful for space applications to protect polymers that are vulnerable to oxidation or degradation by oxygen atoms.

INTRODUCTION

Considerable interest exists in the surface recession (etching) of polymers exposed to ground-state atomic oxygen [O(^{3}P)] in low Earth orbital environment, and some effort has been devoted to seeking protective coatings for space applications (1). Previously, we reported that polybutadienes and related unsaturated hydrocarbon polymers are quite reactive towards O(^{3}P) generated by a radio-frequency glow discharge in O_2, the etch rates being dependent upon polymer structure (2). Since Teflon (polytetrafluoroethylene, PTFE) was reported to be resistant to O(^{3}P)-induced etching in the space environment, it was reasonable to expect that plasma-polymerized tetrafluoroethylene (PPTFE) -- a highly branched and crosslinked polymer (3) with a fluorine/carbon ratio of 1.3-1.4 in contrast to 2.0 for the linear PTFE -- would also be resistant. Indeed, since crosslinked natural rubber had been reported to be much more resistant to O_2 plasma etching than the corresponding uncrosslinked polyisoprene (4), there was the possibility that PPTFE could be even more resistant than PTFE. Thus, if PPTFE films, which are easily deposited on various substrates, could be shown to be especially resistant to O(^{3}P) attack, they might be useful as protective coatings for vulnerable polymers deployed in space. This paper describes the O(^{3}P)-induced etching of PPTFE, PTFE and an ion-beam sputter-deposited PTFE (SPTFE), using ESCA to follow surface structural changes produced in these polymers. Thin films of PPTFE deposited onto *cis*-1,4-polybutadiene (CB) are shown to protect the latter polymer against O(^{3}P) attack.

EXPERIMENTAL

PTFE film (25-μm thick) was obtained from Chemical Fabrics Corp., West Palm Beach, FL. A 533-nm film of SPTFE (on a silicon substrate) was kindly supplied by Bruce A. Banks, NASA Lewis Research Center, Cleveland, OH. Tetrafluoroethylene monomer (TFE) inhibited with *d*-limonene (SCM Specialty Chemicals, Gainesville, FL) was used without further purification to prepare PPTFE. PPTFE films were deposited on the polished surfaces of small square sections (1.42 cm^2) of silicon cut from an Si wafer, using the plasma polymerization reactor described previously (5). The following conditions were used for PPTFE deposition:

Power: 10 W at 13.56 MHz
TFE flow rate: 0.5 cm^3 (STP)/min
Pressure: 76 Pa (0.57 torr), discharge off
Deposition rate: 0.066 nm/s

The thickness of the PPTFE films (on Si), before and after etching, was measured with an ellipsometer. The density, determined from thickness and weight-gain measurements on Si substrates of known area, was

2.3 ± 0.1 g/cm^3. The etching of PPTFE, measured in nm, was expressed as mg/cm^2, after multiplying the thinning (in cm) by the density (in mg/cm^3). The etching of SPTFE was similarly followed by ellipsometry. Weight loss in the PTFE films (disks cut to area of 2.54 cm^2) was measured on an automatic electrobalance and likewise expressed as mg/cm^2.

The O(^{3}P) reactor (2) is shown in Figure 1. The O_2 and O atom flow rates (the latter obtained by titration with NO_2) were 3.9 and 1.4 cm^3 (STP)/min, respectively, and O atom partial pressure was 0.17 torr at the NO_2 titration port. Temperature was measured with a thermocouple having a junction located beneath the sample platform. ESCA analysis was performed, using Teflon tape as a standard, and the binding energies of PPTFE and SPTFE were referenced for the single F_{1s} peak of $>$CF-/-CF$_2$-/-CF$_3$ at 689.2 eV.

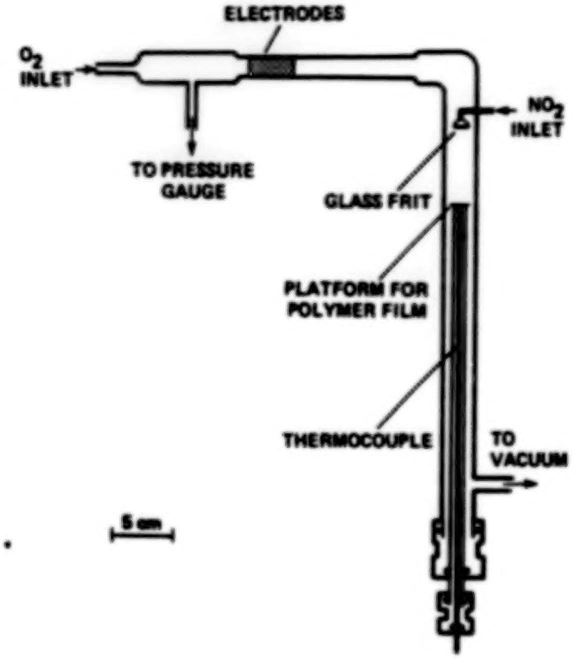


Figure 1. Apparatus for exposure of polymer films to O(^{3}P).

To demonstrate the protectiveness imparted by PPTFE coatings, an Si substrate and a CB film (cast onto a glass cover slip) were simultaneously exposed to TFE for the same length of time in the plasma polymerization reactor. The resulting PPTFE coatings -- assumed to have the same thickness on Si and on CB -- were each subjected to O(^{3}P) for various periods of time. Thickness measurements were made on the former coating until the PPTFE was completely etched away, while weight-loss measurements were made on the latter coating to record the onset and subsequent etching of CB.

RESULTS AND DISCUSSION

In our original paper (2) we presented kinetic plots for O(^{3}P)-induced surface recession, or etching, of PPTFE and PTFE films at temperatures ranging from 22°C to 177°C. In contrast to PTFE, which is stable throughout this temperature range, PPTFE undergoes a thermally induced thinning in the absence of O atoms, which becomes increasingly significant as the temperature exceeds ca. 50°C. For the purposes of this presentation, we shall confine our attention to kinetic data obtained at 22°C; at this temperature, the etch rate for PTFE (5.3 x 10^{-4} mg/cm^2 h) was found to be 8.7 times that of PPTFE (6.1 x 10^{-5} mg/cm^2 h). That PPTFE has a higher resistance than PTFE to O-atom attack at ambient temperature is accounted for on the basis that the former polymer has a crosslinked structure (3) while the latter has a linear structure. SPTFE, on the other hand, had an etch rate at 22°C (1.1 x 10^{-4} mg/cm^2 h) that was 1.8 times that of PPTFE but only 0.21 times that of PTFE. That SPTFE had an etch rate closer to that of PPTFE than to that of PTFE is not surprising since ion-beam sputtering of PTFE involves breakdown of the latter's linear polymer structure and reconstitution elsewhere of a polymer network from the PTFE fragments. Based on ESCA results given below, SPTFE has a structure that is less crosslinked than PPTFE and contains short sequences of CF_2 groups. At any rate, the lower etch rate for SPTFE compared to PTFE reinforces the viewpoint that crosslinking enhances resistance to $O(^3P)$-induced etching of an otherwise linear polymer.

The protectiveness of a very thin PPTFE coating against $O(^3P)$-induced etching at ambient temperature is illustrated in Figure 2 by means of two PPTFE coatings having the same initial thickness (16.8 nm) -- one deposited onto an Si substrate (for following decrease in thickness), the other onto a CB

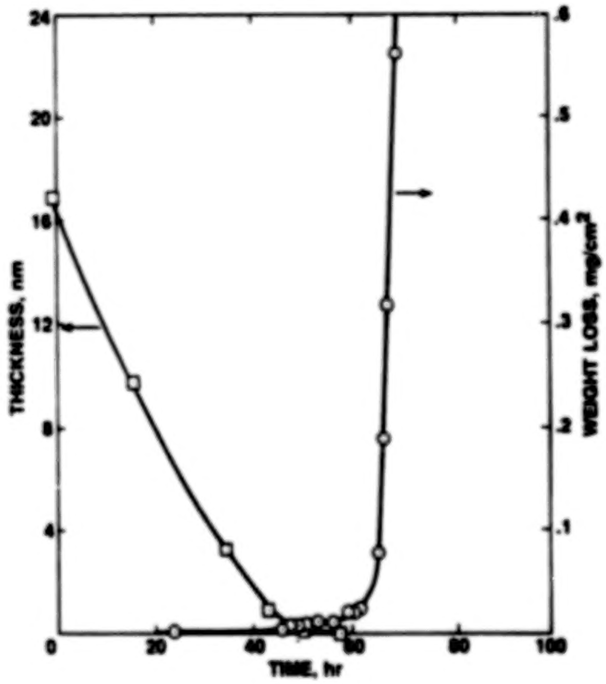
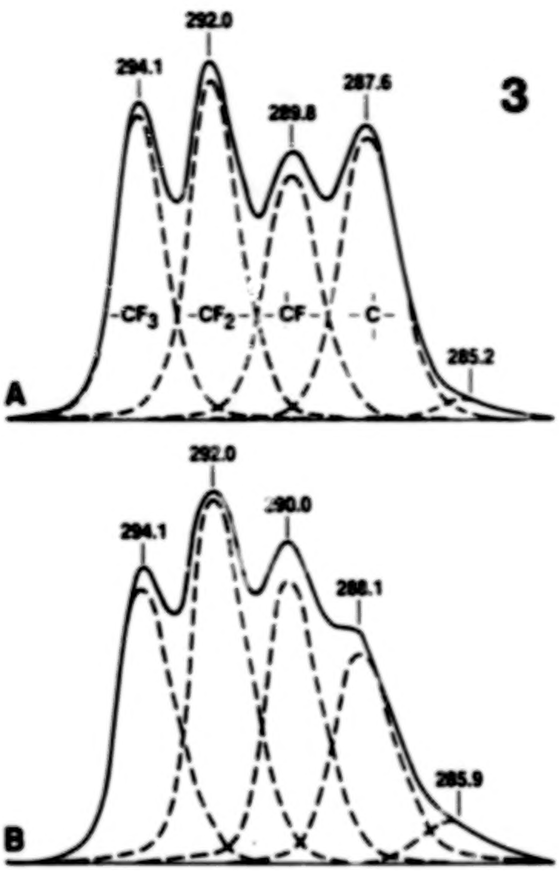


Figure 2. Protective effect imparted by a PPTFE coating against $O(^3P)$ attack at 22°C: (□) thinning of a PPTFE coating on Si substrate; (○) weight loss in a PPTFE coating on a CB film cast on a glass cover slip. Both PPTFE coatings had the same initial thickness (16.8 nm).

film cast on a glass cover slip (for following weight loss). As may be seen, the onset of significant weight loss in the PPTFE/CB sample occurs only after a time interval (ca. 50-60 h) required for the PPTFE coating to be nearly completely etched away (the weight of the PPTFE coating being immeasurable in this sample). This is demonstrated by the fact that the PPTFE/Si sample reaches a negligible thickness (0.2 nm) after this same time interval. In fact, the "breakthrough" point is in good agreement with the predicted value of about 62 h, given an etch rate for PPTFE of 0.27 nm/h ($\approx 6.1 \times 10^{-5}$ mg/cm² h) at 22°C. Moreover, the sharp rise in the weight-loss curve of PPTFE/CB, subsequent to the disappearance of the PPTFE coating, yields an etch rate for CB of 0.18 mg/cm² h, which agrees quite well with the previously reported etch rate of 0.13 mg/cm² h for CB in the same $O(^3P)$ reactor (2). The near constant weight of the PPTFE/CB sample prior to total disappearance of the PPTFE coating indicates that O atoms do not diffuse into PPTFE to any significant extent. At the same time, it implies that the PPTFE coating is *conformal* and free of pinholes.

The C_{1s} and O_{1s} regions of the ESCA spectra of PPTFE films before and after etching at 22°C are shown in Figures 3 and 4. The analytical data are summarized in Table 1, where the estimated accuracy of each deconvoluted peak area (given in parentheses) is ±5% of its given value, while the estimated accuracy of each F/C ratio is ±2% of its given value. Since the F_{1s} regions were unchanged and, moreover, resembled the single intense peak at 689.2 eV ($-CF_x$) in the ESCA spectra of PPTFE (6) or of Teflon, there is no need to show them here. The C_{1s} spectrum of unetched PPTFE is very similar to the spectra of polymers



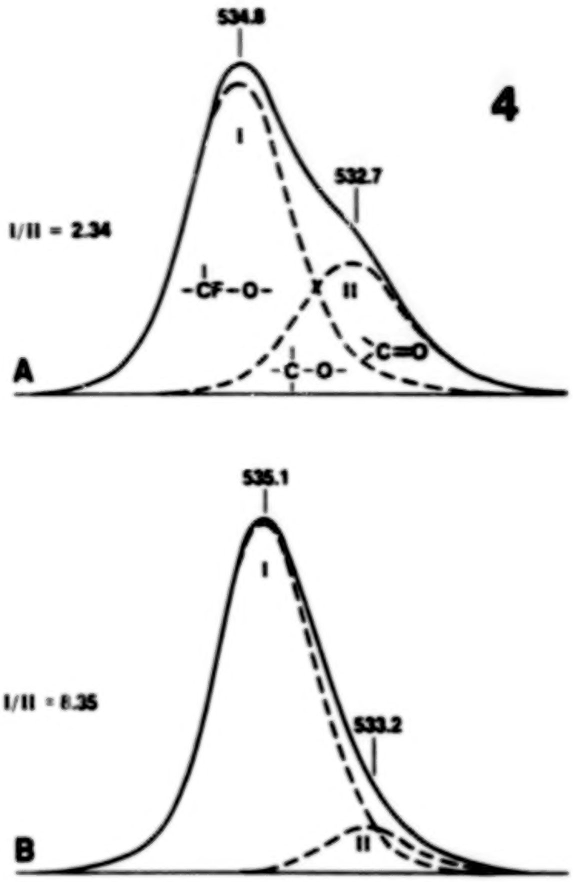


Figure 3. C_{1s} regions of ESCA spectra of PPTFE before (A) and after (B) $O(^3P)$-induced etching. Assignments of peaks under B are the same as those under A, except for additional groups indicated in footnotes of Table 1. Binding energies in this and the following figures are given in eV units.

Figure 4. O_{1s} regions of ESCA spectra of PPTFE before (A) and after (B) $O(^3P)$-induced etching. Spectra are not to scale; area under B corresponds to 8.4 times area under A.

TABLE 1

Atom (and Peak Area) % Distribution in ESCA Spectra of PPTFE[a]

	C_{1s} peaks					O_{1s} peaks		F_{1s} peak	
Condition	$-CF_3$	$-CF_2-$	$-\overset{\vert}{C}F-$	$-\overset{\vert}{\underset{\vert}{C}}-$	$-CH_2-$	$-\overset{\vert}{C}F-O-$	$\geq C=O$	$-CF_2-$	F/C
Unetched	9.9	11.9	9.3	11.3	0.8	0.6	0.2[b]	56.0	1.30
	(23.0	27.6	21.5	26.2	1.7)	(70.1	29.9)	(100.0)	
Etched	9.5[c]	13.1[d]	9.8[e]	8.1[e,f]	1.4[f]	6.0	0.7[b]	51.4	1.23
	(22.6	31.2	23.5	19.3	3.4)	(89.3	10.7)	(100.0)	

[a] Extent of etching at 22°C: ~ 9.5 nm.

[b] May include some $\geq C-O-$.

[c] May include some $-CF_2-O-$.

[d] May include some $-\overset{\vert}{C}F-O-$.

[e] May include some $\geq C=O$.

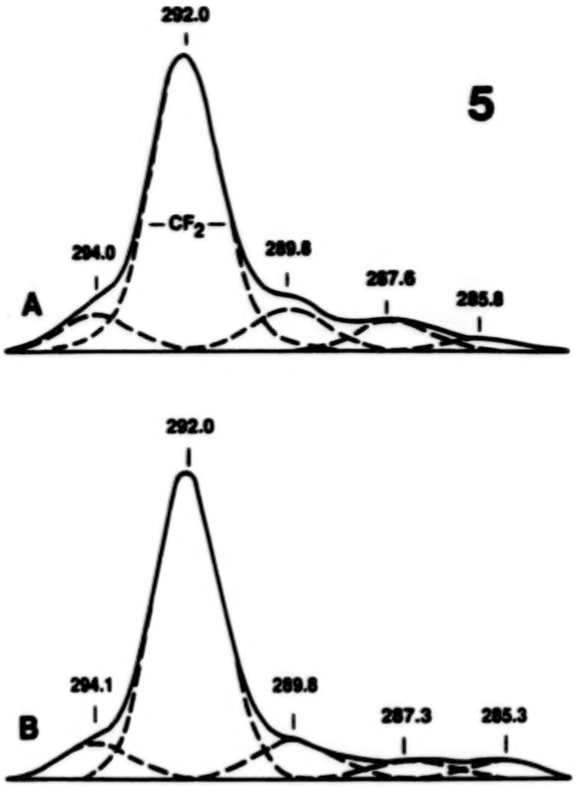
[f] May include some $\geq C-O-$.

obtained by various workers (6,7) from in-glow plasma polymerization of TFE in showing the same four prominent peaks with roughly the same relative intensities. However, these C$_{1s}$ spectra differ from that of a PPTFE produced in nonglow regions (8), which approaches that of ordinary PTFE with its single -CF$_2$- peak. There is universal agreement on the PPTFE assignments (6,7,9) shown in Figure 3(A), and the weak peak around 285 eV signifies a small amount of hydrocarbon contamination (7). To the extent that oxidation of the PPTFE surface occurs on exposure to O(^{3}P), the various C$_{1s}$ peaks in Figure 3(B) contain, besides the functional groups indicated in Figure 3(A), unknown amounts of C atoms attached to O atoms: -CF$_2$-O- (294.1) and >CF-O- (292.0 eV), as well as C atoms in >C=O (288.1-288.9) and >C-O- (286.7-287.5 eV) (see footnotes in Table 1). In addition, the ESCA spectra in Figure 3 are assumed to contain an undetermined amount of carbon atoms attached to double bonds (as in =C<, =CF-, or =CF$_2$ unsaturation (7)), which could comprise some 20% of the total carbon content in the unetched PPTFE (10); however, their exact locations under the various peaks cannot be specified. Hence, the data in Table 1 have only semi-quantitative significance, but they do point to a small decrease in the F/C ratio for PPTFE from 1.30 to 1.23 as a result of oxidative etching. It should be noted that these F/C ratios were determined by dividing the area under the F$_{1s}$ peak by the total area under the five C$_{1s}$ peaks (shown in Fig. 3), and adjusting the number by a sensitivity factor which acounts for the different cross sections for photoionization of F and C atoms. Another method, which involves calculating the F/C ratio from the percentage of each group present in the C$_{1s}$ region and which is less reliable, typically yields F/C ratios higher by 0.2-0.4 than those calculated by the former method. As may be inferred from the data in Table 1, it is difficult to offer a detailed picture of the chemical changes occurring in the surface of the PPTFE film as it undergoes O(^{3}P)-induced etching; however, an essentially stationary state in surface composition is established as fresh surface is being regenerated even as the surface is being etched away.

PTFE, in contrast to PPTFE, showed very little oxygen uptake on exposure to O(^{3}P), despite a more rapid rate of etching at 22°C: the ESCA spectrum for a film of PTFE etched to the extent of 0.4 μm showed only 0.47 atom % O (with an F/C ratio of 1.97), while the spectrum of the unetched PTFE (F/C = 2.00) showed no oxygen at all. The C$_{1s}$ and F$_{1s}$ spectra of PTFE before and after etching appeared unchanged, each exhibiting single peaks at 292.2 and 689.2 eV, respectively; hence the spectra were omitted here. Likewise, the O$_{1s}$ spectrum of etched PTFE, which showed only a single peak at 532.3 eV (>C=O), was also omitted.

The C$_{1s}$ and O$_{1s}$ spectra of SPTFE before and after etching at 22°C are shown in Figures 5 and 6, and the analytical data are summarized in Table 2. Here again, there was no need to show the corresponding F$_{1s}$ spectra for SPTFE, since they each possessed but a single peak at 689.2 eV, in common with the F$_{1s}$ spectra of PTFE and PPTFE. It is obvious from Figure 5 that the C$_{1s}$ spectra of the unetched and etched SPTFE each have the appearance of a composite of the corresponding spectra of PTFE (with a single -CF$_2$- peak at 292.0 eV) and of PPTFE (with its four prominent peaks), the latter two polymers *seeming* to be present in the approximate ratio of 4-5:1. This suggests that our SPTFE is considerably less crosslinked than the highly crosslinked PPTFE.

Since the changes observed in the C$_{1s}$ spectra of PPTFE as a result of O(^{3}P)-induced etching (Fig. 3) are not dramatic, and since no changes were observed in the C$_{1s}$ spectra of PTFE (as mentioned earlier), it is not surprising that the C$_{1s}$ spectral changes in SPTFE (Fig. 5) upon etching are rather slight. However, there is an interesting difference between the O$_{1s}$ spectra of SPTFE and PPTFE: whereas the ratio of areas of the two peaks I and II in Figure 4 increases considerably from 2.34 to 8.35, the corresponding ratio in Figure 6 decreases slightly from 1.53 to 1.39. This difference is likely due to SPTFE having a structure that is some 75-80% like that of PTFE and some 20-25% like that of PPTFE. Thus, for an oxygen uptake in SPTFE of 2.2 atom % (i.e., 2.7-0.5 atom % in Table 2), about 0.9/2.2, or 40%, of the oxidation presumably occurs at PTFE segments to produce the peak at 532.9 eV [Fig. 6(B)], given the prior observation that etched PTFE showed a single O$_{1s}$ peak around 532.3 eV; the remaining 60% of the oxidation can be considered to occur at PPTFE branch points and/or crosslinks to produce the peak at 535.0 eV, given the dominance of the 535.1-eV peak in the O$_{1s}$ spectrum of etched PPTFE [Fig. 4(B)]. Reflecting the mixed PTFE/PPTFE structure of SPTFE, it is worth recalling that the O uptakes in PPTFE, SPTFE, and PTFE -- after sufficient etching time to reach "equilibrium" oxidized surfaces -- were 5.9, 2.2 and ≈ 0.5 atom % O, respectively. Again, our SPTFE, with an initial F/C ratio of 1.73, had an etch rate that was about 2/9 of that of PTFE (with an F/C ratio of 2.0) and about twice that of PPTFE (with an initial F/C ratio of 1.30).



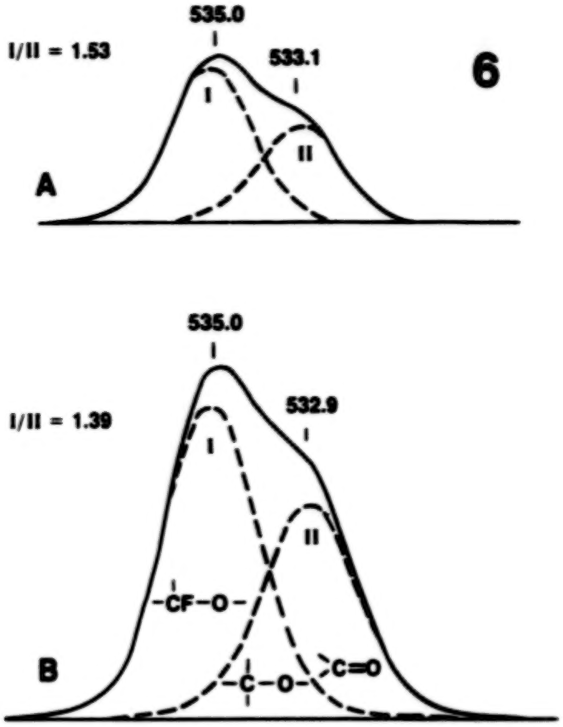


Figure 5. C_{1s} regions of ESCA spectra of SPTFE before (A) and after (B) $O(^3P)$-induced etching. Assignments of peaks under A and B are the same as those indicated in Figures 3(A) and 3(B).

Figure 6. O_{1s} regions of ESCA spectra of SPTFE before (A) and after (B) $O(^3P)$-induced etching. Spectra are not to scale; area under B corresponds to 5.4 times area under A.

TABLE 2

Atom (and Peak Area) % Distribution in ESCA Spectra of SPTFE[a]

	C_{1s} peaks					O_{1s} peaks		F_{1s} peak	
Condition	$-CF_3$	$-CF_2-$	$-\overset{\vert}{C}F-$	$-\overset{\vert}{\underset{\vert}{C}}-$	$-CH_2-$	$-\overset{\vert}{C}F-O-$	$\gt C=O$	$-CF_2-$	F/C
Unetched	3.2	25.4	4.1	2.8	1.0	0.3	0.2	63.0	1.73
	(8.7	69.8	11.4	7.5	2.6)	(60.4	39.6)	(100.0)	
Etched	3.1[b]	26.4[c]	3.6[d]	1.8[d,e]	1.6[e]	1.6	1.1[f]	60.8	1.67
	(8.6	72.1	9.8	5.0	4.5)	(58.1	41.9)	(100.0)	

[a] Extent of etching at 22°C: ~ 48 nm.

[b] May include some $-CF_2-O-$.

[c] May include some $-\overset{\vert}{C}F-O-$.

[d] May include some $\gt C=O$.

[e] May include some $\gt C-O-$.

[f] May include some $\gt C-O-$.

REFERENCES

1. Wydeven, T.; Golub, M. A.; and Lerner, N. R.: Etching of Plasma-Polymerized Tetrafluoroethylene, Polytetrafluoroethylene, and Sputtered Polytetrafluoroethylene Induced by Atomic Oxygen [O(^{3}P)]. J. Appl. Polym. Sci. vol 37, 1989, pp. 3343-3355. See, in particular, references 1-6 cited therein.

2. Golub, M. A.; Lerner, N. R.; and Wydeven, T.: Reactions of Atomic Oxygen [O(^{3}P)] with Polybutadienes and Related Polymers. ACS Symp. Ser. no. 364, 1988, pp. 342-355.

3. Yasuda, H.; and Hsu, T.: Plasma Polymerization Investigated by the Comparison of Hydrocarbons and Perfluorocarbons. Surf. Sci. vol 76, 1978, pp. 232-241.

4. Hansen, R. H.; Pascale, J. V.; De Benedictis, T.; and Rentzepis, P. M.: Effect of Atomic Oxygen on Polymers. J. Polym. Sci. A. vol 3, 1965, pp. 2205-2214.

5. Wydeven, T.: Plasma Polymerized Coating for Polycarbonate: Single Layer, Abrasion Resistant, and Antireflection. Appl. Opt. vol 16, 1977, pp. 717-721; 1138.

6. Dilks, A.; and Kay, E.: Plasma Polymerization of Ethylene and the Series of Fluoroethylenes: Plasma Effluent Mass Spectrometry and ESCA Studies. Macromolecules vol 14, 1981, pp. 855-862.

7. Kaplan, S.; and Dilks, A.: Characterization of Plasma-Polymerized Materials by Modern Spectroscopic Techniques. J. Appl. Polym. Sci.: Appl. Polym. Symp. vol 38, 1984, pp. 105-125.

8. O'Kane, D. F.; and Rice, D. W.: Preparation and Characterization of Glow Discharge Fluorocarbon-type Polymers. J. Macromol. Sci., Chem. vol A10, 1976, pp. 567-577.

9. Inagaki, N.; Nakanishi, T.; and Katsuura, K.: Glow Discharge Polymerizations of Tetrafluoroethylene, Perfluoromethylcyclohexane and Perfluorotoluene Investigated by Infrared Spectroscopy and ESCA. Polym. Bull. vol 9, 1983, pp. 502-506.

10. Perrin, J.; and Dandoloff, R.: Cross-linking and Average Coordination in Plasma-Polymerized Hydro- and Fluorocarbons. J. Non-cryst. Solids vol 86, 1986, pp. 179-189.

N91

24064

UNCLAS

N91-24064

FLAME-RETARDANT COMPOSITE MATERIALS

Demetrius A. Kourtides
NASA Ames Research Center
Moffett Field, CA 94035

ABSTRACT

This paper describes the properties of eight different graphite composite panels fabricated using four different resin matrices and two types of graphite reinforcement. The resign matrices included: VPSP/BMI, a blend of vinylpolystyrylpyridine and bismaleimide; BMI, a bismaleimide; Phenolic, and PSP, a polystyrylpyridine. The graphite fiber used was AS-4 in the form of either tape or fabric. The properties of these composites were compared with epoxy composites. It was determined that the blend of vinylpolystyrylpyridine and bismaleimide (VPSP/BMI) with the graphite tape was the optimum design giving the lowest heat release rate.

INTRODUCTION

Graphite-reinforce composites have potential applications in advanced aircraft because of their weight saving and performance characteristics. In this study, graphite panels were fabricated using fabric or unidirectional tape. The contribution of the resin matrix to the ultimate performance of these composites was studied, with particular emphasis on thermal and flammability properties. Comparisons were made with state-of-the-art aircraft composite made with epoxy resin. These advanced composites have applications as fire-resistant and light weight panels for ceiling, floors, and sidewalks for aircraft, space station, mass transit vehicles and other transportation vehicles.

DISCUSSION

Resin Chemistry

Four types of resin matrices were evaluated: a) VPSP/BMI (Hercules, Inc.) a bismaleimide/vinylpolystyrylpyridine (VPSP) formulation; b) BMI (Technochemie GMBH), a bismaleimide; c) Phenolic (Cyanamid Co.), and d) PSP (Societe Nationale et Poudres Explosifs), a polystyrylpyridine. Graphite composites made from these resin matrices were compared with a composite made with an epoxy resin as a matrix. The chemistry of these resins is shown in Figures 1 and 2 and is described below.

1. Epoxy Resin: The baseline epoxy resin was an amine-cured polyfunctional glycidyl amine-type epoxy resin.

2. VPSP/BMI: This formulation is based on a formulation of bismaleimide (BMI) and a modified VPSP designated as XU71775.01L (Dow Chemical Co.). This resin has the same oligomer backbone (polystyrylpyridine) as VPSP, but possesses different reactive end groups. The chemistry of the VPSP has been described previously (1, 2, 3). The VPSP/BMI formulation contains seven parts by weight BMI and three parts by weight XU71775.01L. Other reactive materials are added to this formulation to allow hot melt prepreging of the resin. This resin was characterized thermally by differential scanning calorimetry (DSC), thermogravimetric analysis (TGA), and exothermicity. The DSC of the resin was measured at 10° C/min in nitrogen. The endothermic peak at 60-140° C is probably due to the evaporation of trace amounts of volatiles. The exothermic cure temperature is at 160-240° C with a cure peak at 211° C. The resin can be cured at 179° C for a longer time. After heating at 179° C for 3 hours, the resin showed no residual cure peak. The exothermic peak at 300-360° C is probably related to the decomposition of the resin. The TGA of the cured pure resin in nitrogen and air at 10° C/min was determined. The resin starts to decompose at about 320° C. The char yield at 800° C is 52.5%. The TGA of epoxy resin is also shown.

3. BMI: This bismaleimide resin is produced by reacting m-maleimidobenzoic acid chloride with an aromatic diaminocompound in the molar proportion of difunctional amine acid halide 1.4:2. The resulting resin consists of a mixture of a bismaleimide and an aminoterminated monoimide.

4. Phenolic: The chemistry of this phenolic resin is not known.

5. PSP: The reaction scheme for this resin is shown in Figure 2. This resin has been described previously in detail in (4).

Composites Description

The resins just mentioned were used to fabricate eight types of composite panel using two types of reinforcements: a) plain-weave woven graphite fabric (A-193, Hercules, Inc.), and b) unidirectional tape graphite fiber (AS-4, Hercules, Inc.). All panels were fabricated using a honeycomb core (HR-10, Hexcel, Inc.). All composites had a film of polyetheretherketone (PEEK, Imperial Chemical Co.) adhered with a silicone adhesive (Dow Corning X3-5815) on one side. The composition of the panels is given in Figure 1. The thickness of the panels varied slightly depending on the number of plies in each panel. Panels constructed with the graphite fabric had one ply on each side, and panels fabricated with the graphite tape had three plies on each side placed at 0°, 90°, and 0° orientation for maximum strength. The processing of the baseline panel consisting of epoxy-glass fabric with a polyvinylfluoride (PVF) film has been described previously in detail (5). Panels type A, B, E, and F were cocured with the honeycomb core without the use of an additional adhesive. On panels type C, D, G, and H, a polyimide adhesive film (FM-34, Cyanamid Co.) was used to bond the face sheets to the honeycomb. The fabrication procedure and the curing schedule for these composites have been described previously (6). The density of the panels is shown in Figure 2.

TEST RESULTS AND ANALYSIS

A broad range of flammability and thermal tests were conducted to characterize the composites. Three basic flammability properties of the materials were measured: a) propensity to burn, or oxygen index (OI); b) smoke emission, and c) heat release.

Oxygen Index

The composite panels were tested by the OI in accordance with ASTM D-2863-77(7). The intent of the OI test method is to determine the relative flammability of plastics by measuring the minimum concentration of oxygen in a slowly rising mixture of oxygen and nitrogen that will just support combustion; i.e., OI is defined as the minimum concentration of oxygen and nitrogen that will just support combustion of a material under the conditions of this method.

The test results shown in Figure 3 were as follows: Baseline epoxy-glass fabric/PVF 34.6; panels type A -- VPSP/BMI/Fabric/PEEK 44.3; B -- VPSP/BMI/Tape/PEEK/45.6; C -- BMI/Fabric/PEEK 35.7; D -- BMI/Tape/PEEK 45.0; E -- Phenolic/Fabric/PEEK 38.8; F -- Phenolic/Tape/PEEK/36.9; and G -- PSP/Fabric/PEEK 40.3.

Panel D-BMI/Tape/PEEK had the highest oxygen indes of all the panels tested, followed by Panel A-VPSP/BMI/Fabric/PEEK. The baseline epoxy-glass fabric/PVF had the lowest oxygen index.

Smoke Emission

The smoke emission characteristics of the panels were determined by heating the composites in the NBS Smoke Chamber and the Ohio State University (OSU) heat-release apparatus (8). In this method, the specimen to be tested is injected into an environmental chamber through which a constant flow of air passes. The specimen's exposure is determined by a radiant heat source adjusted to produce the desired total heat flux of 3.5 W/cm~ on the specimen. The specimen is tested so that the exposed surface is vertical. Combustion is initiated by a piloted ignition. The smoke is measured with a photoelectric tube mounted on top of the apparatus. The smoke density is calculated by integrating the light transmission loss over the length of the run.

Specify optical density,

$$\text{maximum}=D_s=\frac{V_2}{AL}\int_0^t \text{Log}_{10}\ \frac{(100)}{(\text{Plt})}\ \frac{T_o}{T_i}\,dt \quad (1)$$

Ds = specific optical density, t = time V2 = volume of air, 2.4 m^3/min A = area of sample, 232.3 cm2 L = length of light path, 0.93 m Plt = percent of light transmission Ti = inlet temperature To = outlet temperature dT = time interval. The test results are given in Figures 4-6. The baseline epoxy-glass fabric/PVF composite had the highest smoke evolution of all composites tested. Panel type H (PSP/Tape/PEEK) had the lowest smoke evolution. The relative ranking of the composites in terms of increased smoke density at 90 sec is as follows: panels type H, B, D, F, C, A, E, G, and Baseline.

Heat Release

The heat-release rates of the composite panels were determined using the OSU Release Calorimeter (8) using a revised test method (9). In this procedure, the specimen to be tested is injected into the environmental chamber through which a constant flow of air passes. The specimen's exposure is determined by using a calibrated calorimeter on a radiant heat source adjusted to produce the desired total heat flux of 3.5 W/cm^2 on the specimen. The temperature difference between the air entering the environmental chamber and that leaving is monitored by a thermopile having three hot, and three cold, 32-gauge Chromel-Alumel junctions. The hot junctions are spaced across the top of the exhaust stack. The cold junctions are located in the pan below the lower air-distribution plate. Heat-release rates are calculated from the reading of the thermopile output voltage at any instant of time as

$$\text{HRR}=\frac{(V_m-V_b)\times K_h}{0.02323\ m^2}$$

HRR = heat-release rate, kW/m^2
Vm = measured thermopile voltage, mV
Vb = "blank" thermopile voltage test obtained by a run conducted with an empty sample holder assembly
Kh = calibration factor, kW/mV

The integral of the heat-release rate is the total-heat release. According to regulations for aircraft (10), the total-heat release over the first 2 min of sample exposure shall not exceed 65 kW-min/m^2 and the peak-heat-release rate shall not exceed 65 kW/m^2. Figures 7 and 8 give the total heat release of the composite panels when exposed at a heat flux of 3.5 W/cm^2. The peak heat-release rate in kW/m^2 of the panels was: baseline epoxy-glass fabric/PVF 82; panel types A -- VPSP/BMI/Fabric/PEEK 57; B -- VPSP/BMI/Tape/PEEK 51; C -- BMI/Fabric/PEEK 84; D -- BMI/Tape/PEEK 91; E -- Phenolic/Fabric/PEEK 80; F -- Phenolic/Tape/PEEK 83; G -- PSP/Fabric/PEEK 99; and H -- PSP/Tape/PEEK 67.

According to these test results, of the nine panels tested, only panel type B (VPSP/BMI/Tape/PEEK) met the above criteria. The total-heat release of this composite was 62 kW-min/m^2 and the peak-heat release rate was 51 kW/m^2. Panels types A and H, with heat release of 66 and 67 kW/m^2, respectively, were marginal failures. The flexural strength, flatwise tensile strength and peel strength of the composites are shown in Figures 11 and 12.

CONCLUSION

To rank the composites, one should consider all of the materials parameters and assign weight to each specific parameter or measurement. Recent studies (9) have indicated that a low rate of heat release or fuel contribution is one of the most important considerations when using composites in critical applications such as aircraft. Based on these observations, the following conclusions may be drawn from this study:

1. The highest total-heat release and heat-release rates were measured with the baseline epoxy composite panel. This panel also exhibited the highest smoke evolution and lowest oxygen index of all the composites tested.

2. The type B panel (VPSP/BMI/Tape/PEEK) exhibited the lowest heat-release rate and total-heat release. It was the only panel tested which meets the performance criteria of maximum heat-release rate of 65 kW/m^2 and maximum total heat release in 2 min of 65 kW-min/m^2. This composite panel measured the highest oxygen index of all the panels.

3. The lowest smoke evolution was measured in panel type G (PSP/Fabric/PEEK) and panel type D (BMI/Tape/PEEK).

4. All the graphite composites exhibited oxygen indices significantly higher than the baseline panel, indicating that the graphite panels will exhibit lower relative flammability. Composite panels A, B, and D showed the highest indices (44.3, 45.6, and 45.0, respectively), compared to 34.6 for the baseline panel.

5. The aforementioned data indicate that composites fabricated with the VPSP/BMI vinylpolystyrylpyridine/bismaleimide resin exhibited the optimum combination of fire-resistant properties and processing characteristics.

REFERENCES

1. D. A. Kourtides, 56th Mtg. Soc. Rheology, 1, p. 260 (1985).

2. M. T. Hsu, J. A. Parker, T. Chen, and A. H. Heimbuch, 29th National SAMPE Symposium, 21, p. 1034 (1984).

3. U.S. Patent 4526925, "Copolymers of Vinyl polystyrylpyridine or Vinyl Stilbazoles with Bismaleimide" (1985).

4. B. Malasine, 24th National SAMPE Symposium, 24, p. 304 (1979).

5. D. A. Kourtides, W. J. Gilwee, and J. A. Parker, J. Polymer. Eng. and Sci., 19, p. 226 (1979).

6. W. D. Roper, NASA CR-177397 (1986).

7. "Standard Test Method for Measuring the Minimum Oxygen Concentration to Support Candle-like Combustion of Plastics (Oxygen Index)," Amer. Soc. Testing and Materials, ASTM D 2863-77 (1979).

8. "Standard Test Method for Heat and Visible Smoke Release Rates for Materials and Products," Amer. Soc, Testing and Materials. ASTM E 906-83 (1982).

9. R. G. Hill, T. I. Eklund, and C. P. Sarkos, "Aircraft Interior Panel Test Criteria Derived from Full-Scale Fire Tests," DOT/FAA/CT-85/23 (1985).

10. "Improved Flammability Standards for Materials Used in the Interiors of Transport Category Airplane Cabins," Department of Transportation -- F.A.A., Federal Register, Vol. 5, No. 139, p. 26206 (1986).

Graphite Composite Makeup

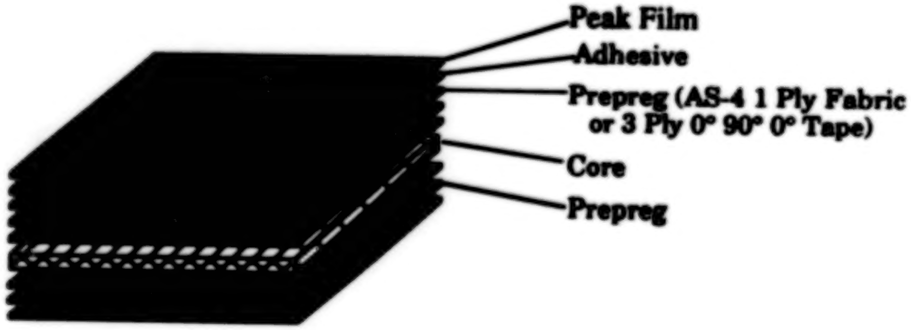

Peak Film
Adhesive
Prepreg (AS-4 1 Ply Fabric or 3 Ply 0° 90° 0° Tape)
Core
Prepreg

Figure 1

Density of Composite Panels
(All fibers are AS-4 Graphite except as noted)

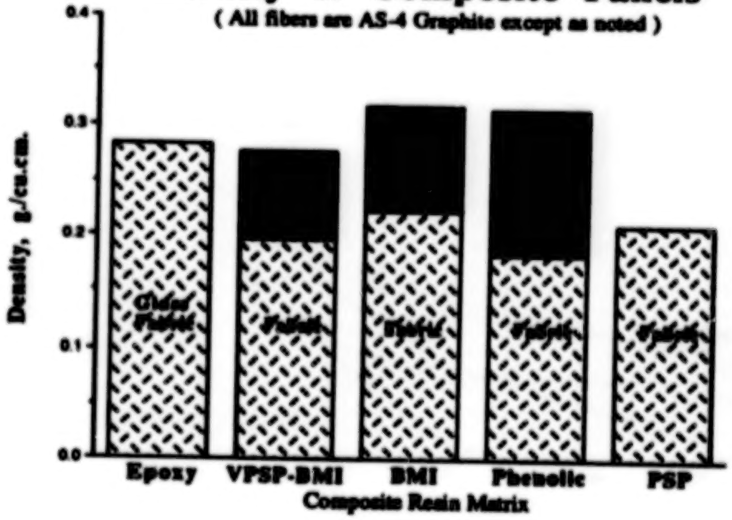


Figure 2

Oxygen Index of Composite Panels
(All fibers are AS-4 Graphite except as noted)

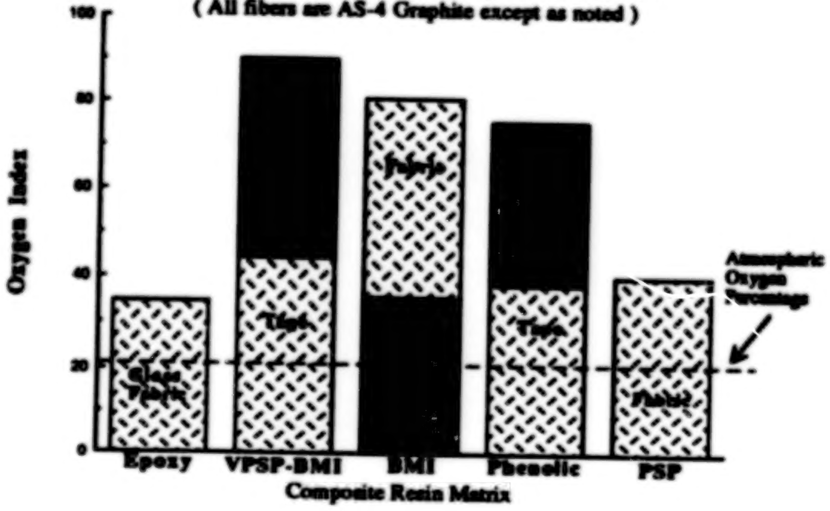


Figure 3

Specific Optical Density of Fabric Composite Panels in the NBS Smoke Chamber
(All fibers are AS-4 Graphite except as noted)

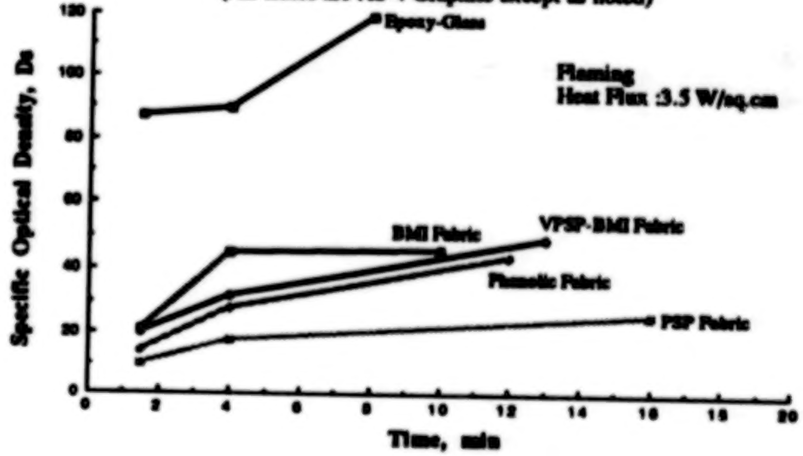


Figure 4

Specific Optical Density of Fabric Composite Panels in the OSU Heat Release Apparatus
(All fibers are AS-4 Graphite except as noted)

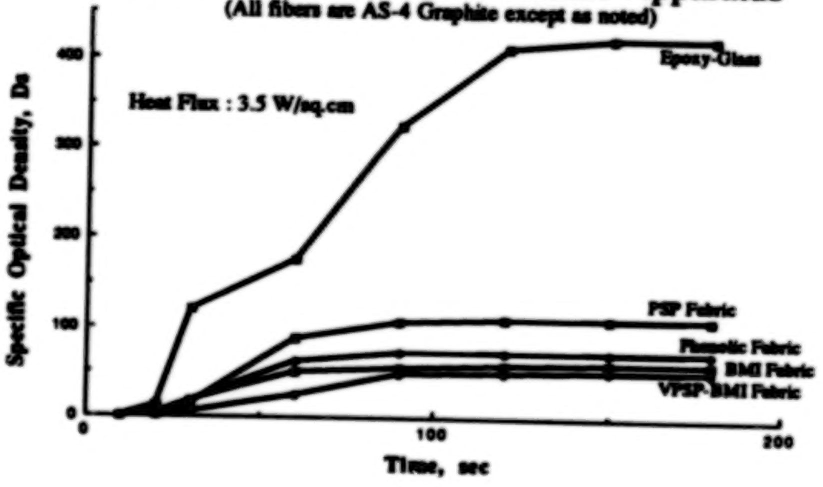


Figure 5

Specific Optical Density of Unidirectional Composite Panels in the OSU Heat Release Apparatus
(All fibers are AS-4 Graphite except as noted)

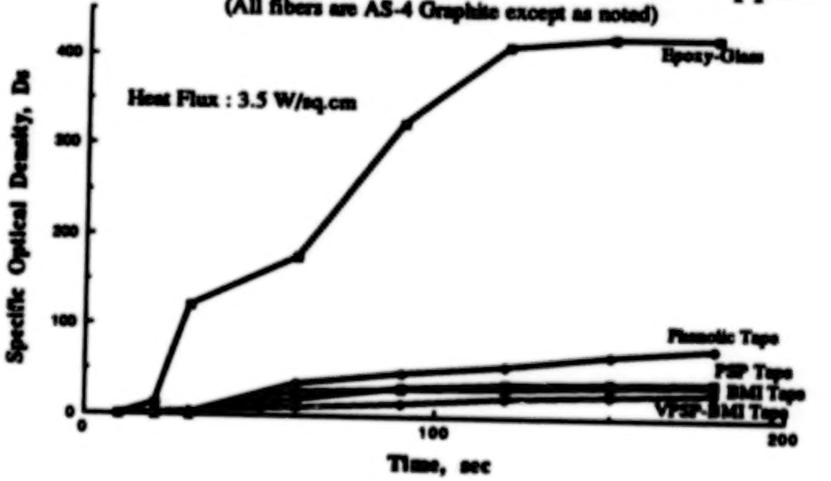


Figure 6

Total Heat Release of Fabric Composite Panels
(All fibers are AS-4 Graphite except as noted)

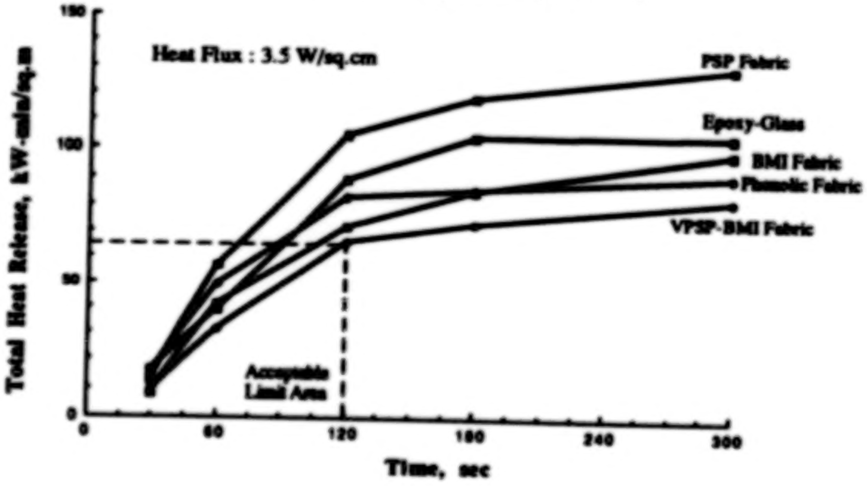


Figure 7

Total Heat Release of Unidirectional Composite Panels
(All fibers are AS-4 Graphite except as noted)

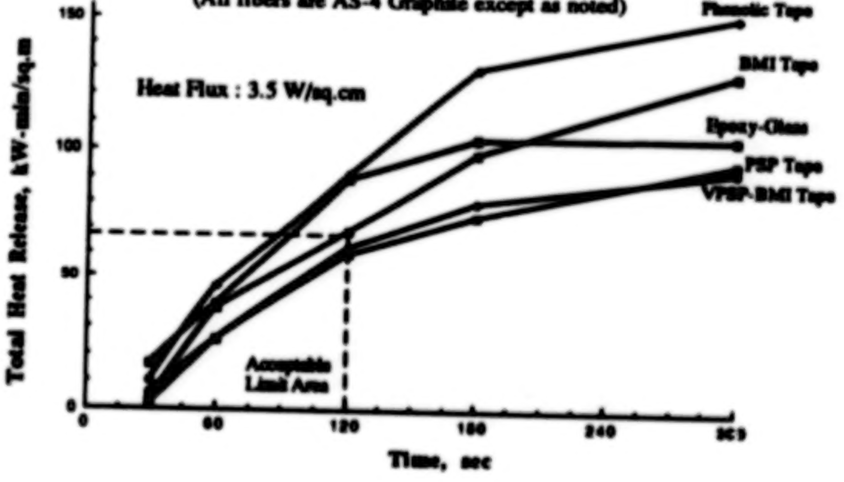


Figure 8

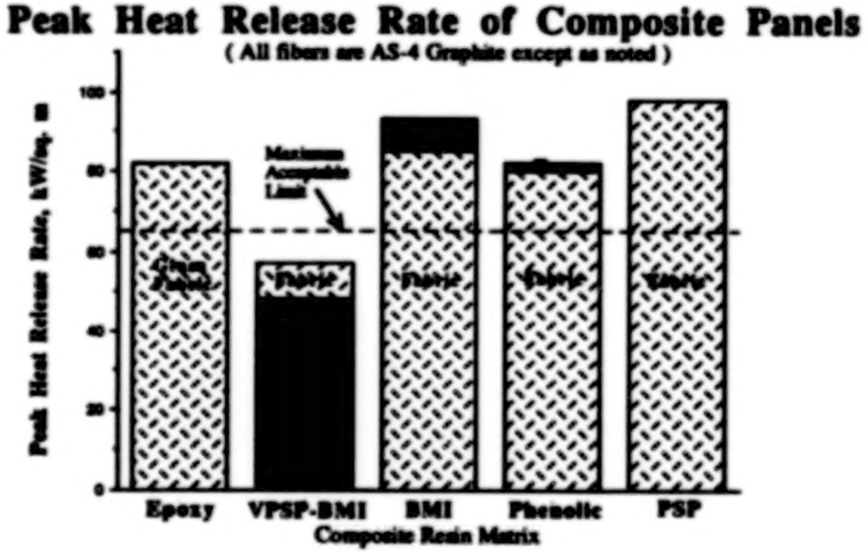


Figure 9

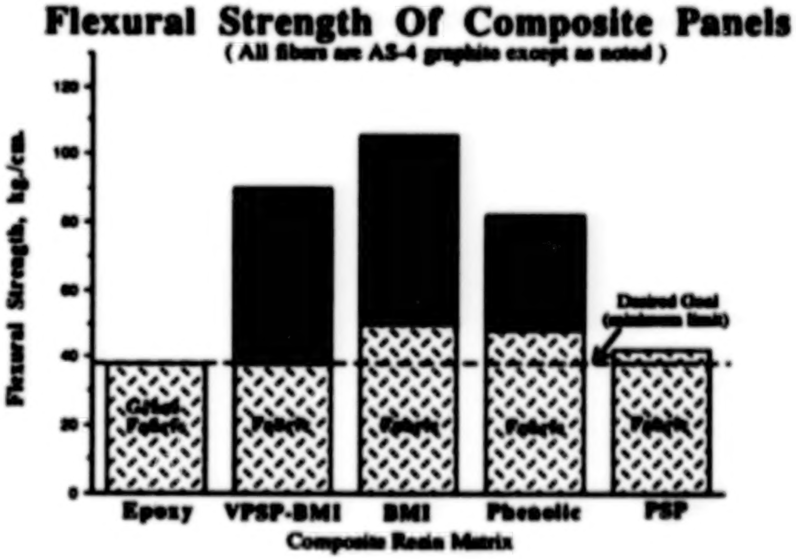


Figure 10

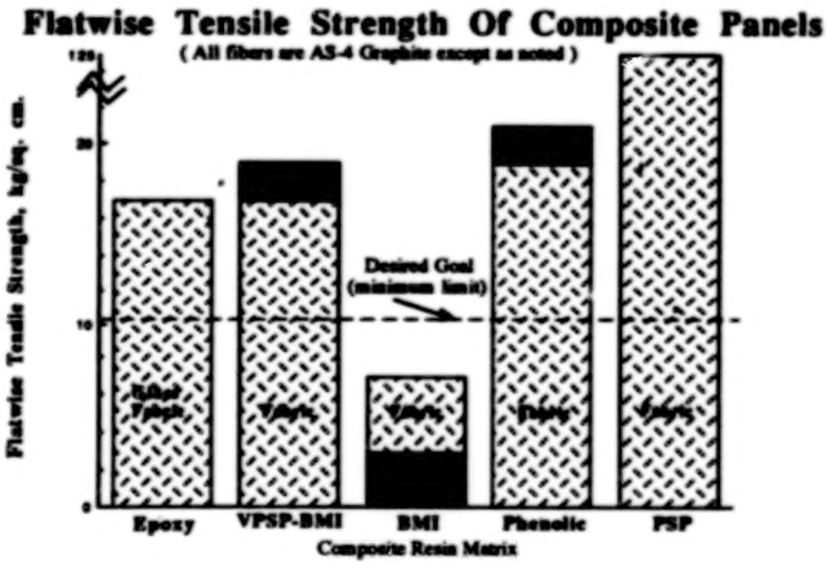


Figure 11

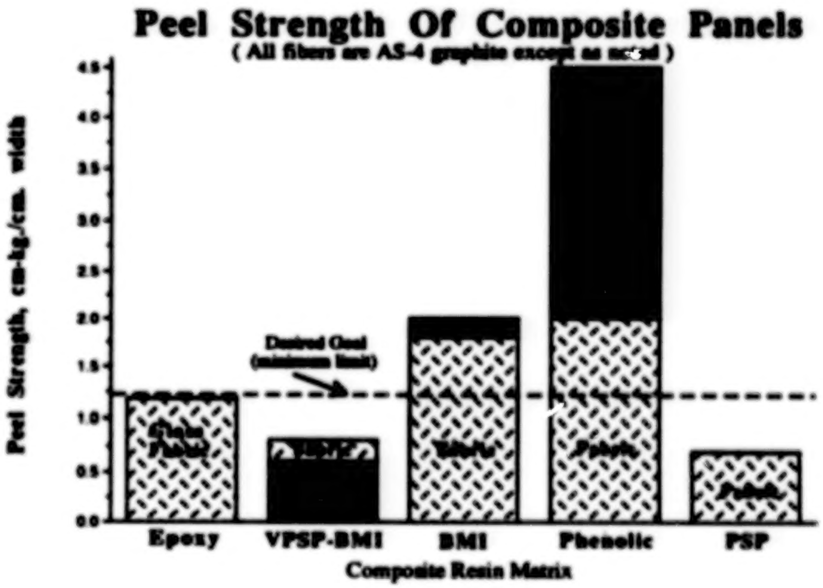


Figure 12

N91

24065

UNCLAS

N91-24065

SUPERPLASTIC FORMING OF AL-LI ALLOYS FOR LIGHTWEIGHT, LOW-COST STRUCTURES

Stephen J. Hales

Analytical Services & Materials
Materials Division, MS 188A
NASA Langley Research Center
Hampton, VA 23665-5225

John A. Wagner

Materials Division, MS 188A
NASA Langley Research Center
Hampton, VA 23665-5225

ABSTRACT

Superplastic forming of advanced aluminum alloys is being evaluated as an approach for fabricating low-cost, light-weight cryogenic propellant tanks. Built-up structure concepts (with inherent reduced scrap rate) are under investigation to offset the additional raw material expenses incurred by using aluminum-lithium alloys. This approach to fabrication offers the potential for significant improvements in both structural efficiency and overall manufacturing costs. Superplasticity is the ability of specially processed material to sustain very large forming strains without failure at elevated temperatures under controlled deformation conditions. It has been demonstrated that superplastic forming technology can be used to fabricate complex structural components in a single operation and increase structural efficiency by as much as 60 percent compared to conventional configurations in skin-stiffened structures. Details involved in the application of this technology to commercial grade superplastic aluminum-lithium material are presented. Included are identification of optimum forming parameters, development of forming procedures and assessment of final part quality in terms of cavitation volume and thickness variation.

INTRODUCTION

Conventional methods for fabricating Al alloy sheet product, such as brake forming or roll forming, usually involve multiple-step cold forming and intermediate heat treating operations. The forming loads required exceed the material yield stress and for high strength alloys the maximum allowable elongation is typically less than 10 percent [1]. This limited ductility restricts the capabilities of conventional forming to relatively simple structural shapes for high strength Al alloys. In comparison, by utilizing the capabilities of superplastic materials, parts can be designed with up to 200 pct. thickness strain and formed at relatively low stresses. This creates versatility in the shapes that can be formed and permits the fabrication of components with more complex geometries [1].

There are many instances in which the increased formability afforded by the superplastic forming (SPF) process has been employed where part complexity has dictated that conventional forming techniques are impractical [2]. Fabrication of sheet structures using SPF technology has permitted the replacement of components consisting of multiple parts, with a single, or greatly reduced number of parts [1]. The technology has also provided additional production flexibility through the capability of fast turnaround for redesigned parts [2]. Overall savings in manufacturing costs have been realized as a consequence of reductions in part/fastener inventories, assembly operations and material scrap rate [3]. The level of automation allowed by the process has reduced the reject rate through decreased reliance on operator skill [4]. The dimensional accuracy of individual components and the repeatability between components has also caused significant reductions in the cost of secondary forming and straightening operations [2].

SPF technology is also a viable alternative where multiple-step forming has proved to be prohibitively expensive [5]. Such instances typically involve production of a restricted number of complex components. Cost savings have been realized through the fabrication of parts in a single forming operation and the elimination of multiple pressing and heat treatment operations [5]. As a result of the requirement for sets of precision tooling with moving and mated die components, the non-recurring tooling costs are high in conventional forming operations [6]. In order to be cost effective, the production volume is necessarily large to offset these expenses. By comparison, the cost of SPF tooling is much lower, as a result of the requirement for one simple cavity die [1]. However, the dedicated control equipment necessary and the relatively low throughput of forming machines combine to inflate costs [7]. Therefore, the cost balance for SPF favors low to moderate volume, where the lower tooling costs are offset by the higher costs of longer production cycle times [7]. The specific volume for which SPF is cost effective is dependent on component complexity and the number of conventional forming steps eliminated [5].

It is clear that for cost effective application of SPF technology there are many factors to be considered which require an understanding of the flexibilities and limitations of the basic process. Therefore, it is the goal of this paper to introduce engineers outside the aerospace industry to the advantages of SPF as an alternate forming technique. The intention is to familiarize readers with the process itself and also illustrate in cursory detail how it may be applied.

BACKGROUND

Superplasticity is used to describe the exceptional elevated temperature ductility exhibited by certain materials when deformed under specific conditions. During uniaxial deformation, tensile elongations of 500-1000% are typical, and at flow stress levels well below ambient temperature yield stress. The primary microstructural requirement is an equiaxed grain structure of the order of 10 μm, which is stable at the SPF temperature of the material [8]. The fine grain size can exist in the material prior to SPF or be developed during the SPF process. In both cases, the refined grain structure is promoted by controlled thermomechanical treatments during the processing leading to sheet product. The microstructural stability required for superplastic deformation is usually imparted by a uniform dispersion of fine particles within the material which inhibit grain coarsening [8].

Superplastic behavior is extremely temperature and strain rate sensitive which dictates that there are considerable constraints on the deformation conditions for achieving large elongations. Typically, temperatures in the range of 70-90 percent of the alloy melting point and controlled strain rates in the range of 10^{-4} to 10^{-2} sec^{-1} are required [8]. Within this 'window', the resistance to localized necking is high which results in uniform thinning of the material and large strains to failure. The predominant mode of deformation during superplastic flow is grain boundary sliding which can create microvoids at grain boundary triple points at high strain levels [9]. The formation of these voids, known collectively as cavitation, can be suppressed by the application of a hydrostatic stress during, or following, deformation.

SPF OF STRUCTURAL COMPONENTS

During SPF of actual components, the inherent low forming stresses allow the use of gas pressure as the deformation medium, rather than hydraulic or mechanical driven tools [7]. The most common form of SPF involves biaxial forming of sheet material into female dies containing male inserts in a variety of configurations. A schematic cross-section of the facility constructed at NASA-Langley Research Center (LaRC) for SPF of structural components is presented in Figure 1. The apparatus, which constitutes a pressure vessel, consists of upper and lower steel platens with a gasket area designed into the periphery. The assembly is mounted in a hydraulic press, such that when a compressive load is applied a gas-tight seal is created along the perimeter of the superplastic sheet. The SPF tooling recessed into the platens is composed of a female die box within which removable male inserts are located. The appropriate forming temperature is achieved via resistance heaters located within ceramic blocks attached to the outer surface of the platens. The entire assembly is insulated against heat loss and the tooling is coated with a release agent to facilitate part removal following forming. The forming stress required for SPF is generated by argon gas pressure which is introduced through a manifold using a microprocessor-controlled regulator. The gas is admitted at the rate specified for forming at the constant strain rate selected from prior parametric evaluation of the uniaxial superplastic behavior of the material.

Pressure-time profiles are determined primarily through geometrical consideration of the superplastic sheet at the various stages of forming in conjunction with material flow characteristics [10]. The initial stage of deformation constitutes bulge formation, with the highest rate of sheet thinning occurring at the tip of the forming dome [6]. On die contact, thinning is inhibited by frictional effects and continued thinning during the intermediate stages is restricted to areas of the sheet not in contact with the die. The final stage of the process is when the deforming sheet conforms to the contours of the die [7]. Even though forming profiles are specific to component geometry, the cycle time is primarily a function of the maximum strain and is therefore independent of the size of the sheet being formed. During SPF of complex shapes or multiple parts, the fastest forming element of the overall design can be formed at the optimum strain rate with the remaining elements forming at a lower local strain rate [2].

The rate of gas pressurization is designed to maintain the strain rate for optimum superplastic response as the thickness and geometry of the deforming sheet changes. The thickness distribution in finished parts can be controlled through judicious selection of tool geometry and careful control of process parameters. The thickness of the final component tends to be at a minimum in those locations corresponding to last points of contact of the deforming sheet with the SPF die. Thus, the width-to-depth (aspect) ratio of the female cavity governs the total thickness strain and choice of corner radii for the male insert controls the extent of localized thinning [6]. The type of release agent employed for part removal determines the level of friction between the deforming sheet and the die. This controls the amount of sheet thinning following die contact and ultimately the thickness gradient across the part [6].

Some generic examples of the application of SPF technology to the manufacture of aerospace components are illustrated in Table 1 [2]. The most appropriate applications for this technology have proved to be deep shapes with compound curvature or corrugated panels with complex details [1]. SPF has allowed the fabrication of monolithic structures that are not only more cost effective, but also often more structurally efficient. It is of particular significance to this study that the technology has been extended to include the fabrication of large, integrally stiffened, sheet structures [3]. In previous work at NASA-LaRC on SPF of titanium alloys, it has been shown that the compressive load-bearing capacity of skin-stiffened panels can be increased by as much as 60 percent by using SPF stiffeners with unique web configurations [11,12].

A characteristic of SPF components is the excellent repeatability of part dimensions which yields tolerances equivalent to machining [2]. The structural performance of components tends to benefit from the low variability in shape and thickness from one component to another. From the perspective of structural efficiency, it is desirable to minimize sheet thickness while controlling the thickness variations inherent in SPF components. The capability of the SPF process for tighter thickness tolerances compared to conventional forming techniques allows designs to incorporate thinner gages leading to more efficient structures [1]. An additional advantage, associated with using single surface tooling, is that sheets of different starting thicknesses may be formed to identical dimensions. This creates production flexibility for satisfying specific structural strength, weight and cost parameters [5].

A major emphasis in the development of the SPF process has been concerned with minimizing the volume of cavitation in finished components [4]. Usually, cavitation is effectively suppressed during SPF by the application of a superimposed back pressure, but hot isostatic pressing (HIP'ing) following SPF has also been employed [4]. The level of back pressure required to attain a total void volume of less than 1 percent is typically of the order of the uniaxial flow stress of the material being formed [13]. Prior investigations at NASA-LaRC have demonstrated that cavitation can also be reduced by using a post-forming pressure cycle. This consisted of a specified holding time at the maximum pressure of the equipment at the conclusion of forming [14].

ALUMINUM-LITHIUM ALLOYS

A large development in the aerospace industry at present is concerned with the potential applications of aluminum-lithium (Al-Li) based alloys. The emergence of these alloys, which typically possess 10-20 percent higher specific properties than existing Al alloys, heralds the expansion of SPF technology into primary structure applications [4]. Success with conventional fabrication of sheet product has been limited because Al-Li alloys tend to develop surface shear steps and have an increased propensity for cracking [15]. This has been attributed to prominent microstructural directionality, which also results in mechanical property anisotropy following fabrication. The culmination of this effect is that forming operations must consider the rolling direction in the starting sheet. As a consequence, the scrap rate is necessarily increased to compensate for the directionality in material behavior.

The limitation in formability of Al-Li alloys, typically to radii greater than twice the thickness, has also precluded the fabrication of intricate structural shapes by conventional means [15]. However, since Al-Li alloys can be thermomechanically processed into superplastic sheet, SPF offers the potential to produce complex structural shapes from these materials. Much of the work concerning superplastic versions of these alloys has tended to involve laboratory-scale material, such that the existing database on commercially available material is limited at present [16,17]. The microstructure of superplastic Al-Li alloys is unrecrystallized in the thermomechanically processed condition and relies on conversion into a refined grain structure with the accumulation of strain during SPF [18]. Historically, one of the inconsistencies associated with superplastic materials has been the inability to meet the microstructural prerequisites during scale-up to commercial-sized ingots. Material characteristics have tended to vary from lot to lot, such that it has proved necessary to check the superplastic properties of each heat. Published literature concerned with the manufacture of superplastic Al-Li alloys and the practical aspects of SPF has also been limited because of the commercial sensitivity of the information [19].

The commercial Al-Li alloys available at present are 2090 produced by ALCOA, 8090 produced by British Alcan, and X2095 (formerly Weldalite™049), produced by Reynolds. Information on the SPF behavior of these alloys is available from the vendors, but much less data exist on the effect of SPF on mechanical properties [18]. Most data have been generated from uniaxially deformed sheet rather than from biaxially formed material. The typical post-SPF properties of these alloys in the T6 (peak aged) condition is compared with 2219 in the T8 (cold stretched + peak aged) condition in Table 2. It should be noted that as with many superplastic materials there is reduction in the strength of Al-Li alloys in the post-SPF condition. However, for the purposes of this investigation, it can be seen that the mechanical properties of post-SPF materials are comparable with those of the current 2219 cryogenic tank material. It should be noted that a yield strength of 87 ksi, with an elongation of 6 percent, for the X2095 material is impressive relative to other Al alloys in any condition.

SPF OF AL-LI COMPONENTS FOR CRYOGENIC TANKS

One activity at NASA-LaRC is focussed on the use of built-up sheet metal structures in the manufacture of cryogenic propellant tanks. The program advocates SPF and Resistance Spot Welding (RSW) of Al-Li alloys for low-cost fabrication of skin-stiffened structures with reduced structural weight. The Space Shuttle External Tank is currently fabricated from alloy 2219 plate with up to 2" starting thickness. The material is machined into integrally T-stiffened panels and formed to the outer radius of the tank prior to assembly. As indicated in Figure 2, panel fabrication by this route results in a scrap rate of approximately 90 percent. It is envisioned that the built-up structure approach using SPF technology will reduce the scrap rate to approximately 15 percent. The decrease in the material buy-to-fly ratio will allow the use of the more expensive Al-Li alloys and exploitation of improved mechanical properties.

The stiffener configurations considered in this investigation are illustrated schematically in Figure 3 and are compared with the existing integrally T-stiffened design. Since thin sheet structures are generally considered to be stiffness critical from the design standpoint [18], the superplastic versions of the Al-Li alloys 8090, 2090 and X2095, which possess high specific modulus, were considered prime candidates for this application. In order to assess the superplastic behavior of these materials, parametric studies were conducted over a range of temperatures and constant strain rates using uniaxial tension tests. The data presented in Table 3 are a summation of the optimum parameters identified in this investigation for SPF of the three alloys.

For SPF of the stiffener components, gas pressurization cycles were established based on this uniaxial tensile data and consideration of the die geometry. The plain hat configuration had the simplest geometry and was used for the purposes of the analysis summarized in Figure 4 for 8090. The forming cycle was divided into the three distinct stages outlined and the geometry of the unsupported areas of the deforming sheet was approximated by either a sphere or a combination of a sphere and a cylinder. For each stage, the pressure required was calculated via conventional shell theory using instantaneous sheet thickness, radius of curvature and flow stress [10]. The corresponding forming times required were estimated by dividing the incremental true linear strain by the strain rate selected. Interrupted forming cycles were used to determine the extent of forming and revise the predicted cycle where necessary. Using this empirical approach, the durations of the three stages of forming were established and compiled to generate the complete forming profiles [14]. Modifications to the initial forming profile, die geometry and forming procedures were also made to improve final part quality and reduce total forming time. This permitted pressure-time profiles to be developed which maintained a constant rate of deformation corresponding to the strain rate for optimum superplastic formability.

Assessment of the level of back pressure and post-forming pressure required to suppress cavitation was conducted prior to establishing the pressure-time profiles for the forming of actual parts. The results are summarized in Figure 5 for 8090 and X2095, which shows the volume of cavitation in simple 'loaf pans' as a function of the applied hydrostatic pressure normalized with respect to flow stress. As mentioned earlier, the level of hydrostatic pressure required is usually of the order of the flow stress [13]. In order to reduce the total volume of cavitation to less than 1 percent, it is apparent that a pressure somewhat less than the flow stress can be employed to minimize cavitation, particularly if a post-forming pressure cycle is employed. This has significant implications for SPF in facilities with limited pressure capability because it allows for higher differential forming pressures. In this investigation, the maximum inert gas forming pressure, combined with the superimposed back pressure to suppress cavitation could not exceed the 500 psi limitation of the SPF facility. An example of a final forming profile, including superimposed back pressure and a post-forming pressure cycle, is shown in Figure 6 for SPF of 8090 material.

Examination of the initial parts formed revealed localized thinning adjacent to the intersection of the cap with the web. This 'eyebrow' effect was considered undesirable from the perspective of structural performance [11]. The corner radius, which had not been specified during die manufacture, was increased to discourage adjacent thinning, but the effect persisted in subsequent parts [14]. It was concluded upon further examination that deformation of the material in the cap area had apparently continued on contact with the die. Based on prior experience with SPF of titanium alloys, the tooling had been coated with a boron nitride slurry which may be classified as a die lubricant. In an attempt to reduce localized thinning, the SPF cycle was repeated following replacement with a yttrium oxide slurry. Yttrium oxide is considered to be a release agent, rather than a lubricant, since experience has shown that much greater friction is created between the deforming sheet and the die [6]. As a result of this procedure modification, the eyebrow effect was eliminated and the final parts conformed closely to the dimensions of the tooling.

Subsequently, the SPF process was applied to the more complicated stiffener configurations and to the other two Al-Li alloys. The net result of this investigation was that demonstration components of each alloy in each configuration were successfully fabricated using SPF. Components were easily removed from the tooling and were formed with good dimensional conformity to the die. Figure 7 shows examples of the SPF plain web stiffener components fabricated from the three alloys assessed in this study. The parts have been heat treated to peak strength and only partially trimmed for demonstration purposes. Following further trimming these were resistance spot welded to skin material to produce buckling panels for evaluation of relative structural efficiency. Compression panel testing and assessment of the mechanical properties of the materials at ambient and cryogenic temperatures is the focus of current investigations.

SUMMARY

SPF has the potential to significantly reduce manufacturing costs in instances where the requirement is for low to moderate volume production. The technology readily lends itself to the fabrication of sheet metal components which are difficult or impossible to produce using conventional methods. The potential improvement in maintainability and reliability resulting from decreased part and fastener count is particularly attractive. As demonstrated in this investigation, the large forming strains permitted by SPF have resulted in the fabrication of skin-stiffened panels with complex configurations. It is anticipated that employing SPF and Al-Li alloys in the fabrication of built-up cryogenic tank structures will result in weight and cost savings. The reductions in systems costs, through lower density, and manufacturing costs, through lower scrap rate and reduced machining time, are expected to be appreciable. The use of Al-Li alloys is recognized as being restricted to applications where structural weight is critical. However, the cost saving benefits of using SPF technology outlined can be applicable to industries engaged in sheet metal forming operations.

REFERENCES

1. Comley, P.N.: Putting Parts onto Planes - SPF Comes of Age. *Superplasticity in Aerospace*, Edited by Heikkenen, H.C., and McNelley, T.R., The Metallurgical Society of AIME, 1988, pp. 361-370.

2. Stephen, D.: Designing for Superplastic Alloys. *Superplasticity*, North Atlantic Treaty Organization, AGARD-LS-168, 1989, pp. 7.1-7.37.

3. Williamson, J.R.: Aerospace Applications of Superplastic Sheet Forming. *Superplastic Forming of Structural Alloys*, Edited by Paton, N.E., and Hamilton, C.H., The Metallurgical Society of AIME, 1982, pp. 291-306.

4. Barnes, A.J.: Advances in Superplastic Aluminum Forming. *Superplasticity in Aerospace*, Edited by Heikkenen, H.C., and McNelley, T.R., The Metallurgical Society of AIME, 1988, pp.301-313.

5. Sawle, R.: Commercial Applications of Superplastic Sheet Forming. ***Superplastic Forming of Structural Alloys***, Edited by Paton, N.E., and Hamilton, C.H., The Metallurgical Society of AIME, 1982, pp. 307-317.

6. Hamilton, C.H.: Superplastic Sheet Forming. ***Superplasticity***, North Atlantic Treaty Organization, AGARD-LS-168, 1989, pp. 2.1-2.23.

7. Laycock, D.B.: Superplastic Forming of Sheet Metal. ***Superplastic Forming of Structural Alloys***, Edited by Paton, N.E., and Hamilton, C.H., The Metallurgical Society of AIME, 1982, pp.257-271.

8. Pilling, J.; and Ridley, N.: ***Superplasticity in Crystalline Solids***. The Institute of Metals (London), Camelot Press, 1989.

9. Gifkins, R.C.: Mechanisms for Superplasticity. ***Superplastic Forming of Structural Alloys***, Edited by Paton, N.E., and Hamilton, C.H., The Metallurgical Society of AIME, 1982, pp. 3-26.

10. Ghosh, A.K.; and Hamilton, C.H.: Superplastic Forming of a Long Rectangular Box Section - Analysis and Experiment. ***Process Modeling - Fundamentals and Applications to Metals***, American Society for Metals, 1980, pp. 303-331.

11. Davis, R.C.; Mills, C.T.; Prabhakaran, P.; and Jackson, C.: Strucural Efficiency of Corrugated Compression Panels with Curved Caps and Beaded Webs. NASA Technical Paper No. 2272, 1984.

12. Davis, R.C.; Royster, D.M.; and Bales, T.T.: Superplastically Formed Titanium Hat-Stiffened Panels. NASA Techbriefs, Vol. 14, No. 1, 1990, pp. 63-64.

13. Ridley, N.: Cavitation and Superplasticity. ***Superplasticity***, North Atlantic Treaty Organization, AGARD-LS-168, 1989, pp. 4.1-4.14.

14. Hales, S.J.; Bales, T.T.; James, W.F.; and Shinn, J.M.: Fabrication of Structural Components from Commercial Aluminum Alloys using Superplastic Forming. ***Superplasticity in Aerospace II***, Edited by McNelley, T.R., and Heikkenen, H.C., The Metallurgical Society of AIME, 1990.

15. Quist, W.E.; Hyman, M.; Gane, D.H.; and Badger, D.V.: Forming Characteristics of Al-Li Alloys. ***Aluminum-Lithium Alloys - Design, Development and Application Update***, Edited by Kar, R.J., Agrawal, S.P., and Quist, W.E., ASM International, 1988, pp. 401-452.

16. Reynolds, M.J.; Henshall, C.A.; and Wadsworth, J.: Superplastic Forming Characteristics and Properties of Aluminum-Lithium Sheet Alloys. ***Aluminum-Lithium Alloys - Design, Development and Application Update***, Edited by Kar, R.J., Agrawal, S.P., and Quist, W.E., ASM International, 1988, pp. 357-399.

17. Grimes, R.; and Butler, R.G.: The Forming Behavior of Commercially Available Superplastic Aluminum Alloys. ***Superplasticity in Aerospace***, Edited by Heikkenen, H.C., and McNelley, T.R., The Metallurgical Society of AIME, 1988, pp. 97-113.

18. Partridge, P.G.; McDarmaid, D.S.; Bottomly, I.; and Common, D.: The Mechanical Properties of Superplastically Formed Titanium and Aluminum Alloys. ***Superplasticity***, North Atlantic Treaty Organization, AGARD-LS-168, 1989, pp. 6.1-6.23.

19. Grimes, R.: The Manufacture of Superplastic Alloys. ***Superplasticity***, North Atlantic Treaty Organization, AGARD-LS-168, 1989, pp. 8.1-8.16.

Table 1. Typical Aerospace Applications of SPF.

Application:	Examples:
Stability Designed Structures	Ribs Frames Beams Compression Struts
Complex Multi-Element Sheet Components	Panels Mounting Brackets Supports
Complex Envelopes (SPF/DB)	Ducting Tanks Vessels

Table 2. Al-Li Alloy Post-SPF Properties Compared with Al 2219.

Property:	2219-T87	8090-T6	2090-T6	X2095-T6
Density (lbs/in.^{3})	0.102	0.092	0.094	0.098
Modulus (x 10^3 ksi)	10.6	11.3	11.5	11.3
Tensile strength (ksi)	63	60	61	90
Yield strength (ksi)	52	45	47	87
Elongation (pct)	6	6	8	6

Table 3. Optimum SPF Parameters from Uniaxial Data.

Parameter:	8090	2090	X2095
Temperature (°F)	985	950	925
Strain Rate (x 10^{-4} s^{-1})	2.5	5.0	6.0
Failure Strain (pct)	510	775	810
Flow Stress (ksi)	0.4	0.6	0.8

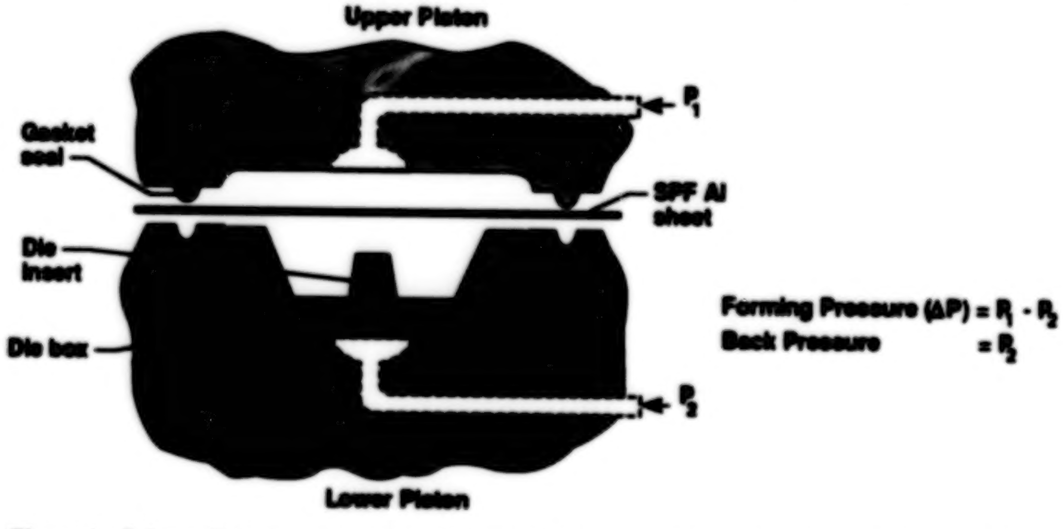

Upper Platen
P_1
Gasket seal
SPF Al sheet
Die insert
Die box
Forming Pressure (ΔP) = P_1 - P_2
Back Pressure = P_2
P_2
Lower Platen

Figure 1. Schematic cut-away representation of the cross-section of the SPF facility used for fabricating structural components. The resistance-heated platen assembly is located within a hydraulic press.

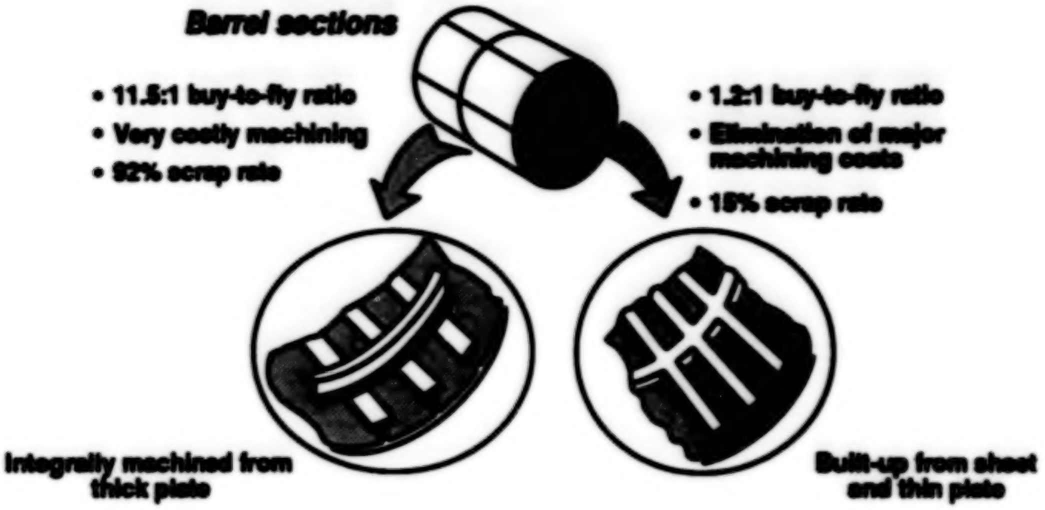

Barrel sections
• 11.5:1 buy-to-fly ratio
• Very costly machining
• 92% scrap rate
• 1.2:1 buy-to-fly ratio
• Elimination of major machining costs
• 15% scrap rate
Integrally machined from thick plate
Built-up from sheet and thin plate

Figure 2. Factors highlighting the potential cost reductions associated with fabricating barrel sections for cryogenic propellant tanks using the built-up structure approach.

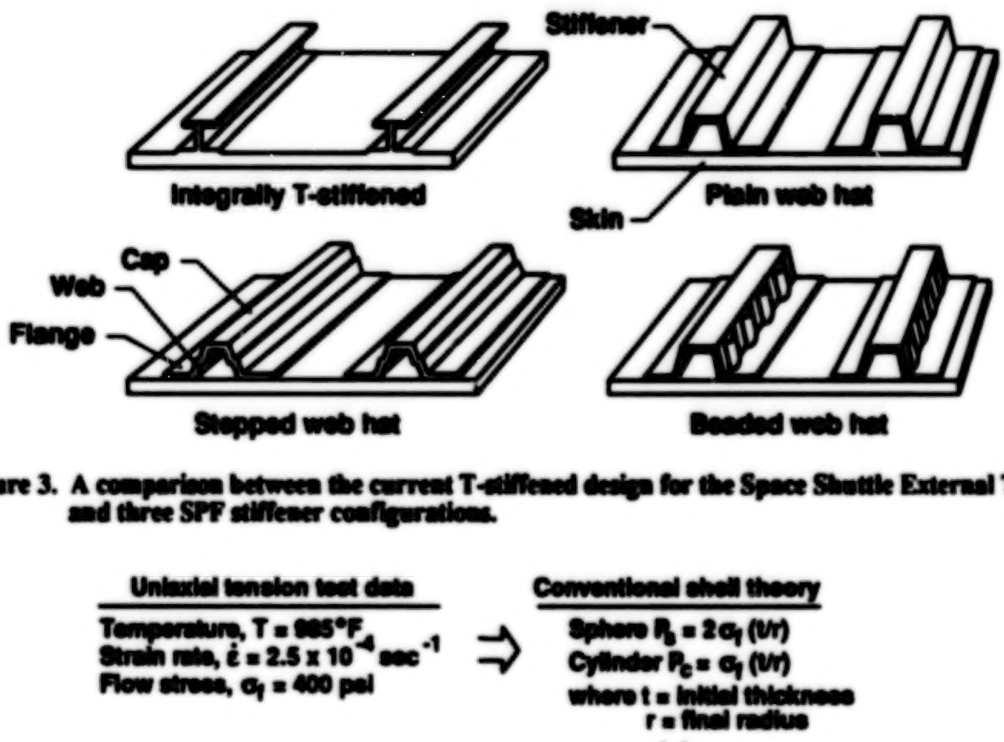

Stiffener
Integrally T-stiffened
Plain web hat
Skin
Cap
Web
Flange
Stepped web hat
Beaded web hat

Figure 3. A comparison between the current T-stiffened design for the Space Shuttle External Tank and three SPF stiffener configurations.

Uniaxial tension test data

Temperature, $T = 985°F$

Strain rate, $\dot{\varepsilon} = 2.5 \times 10^{-4}\ \text{sec}^{-1}$

Flow stress, $\sigma_f = 400$ psi

$\Rightarrow$

Conventional shell theory

Sphere $P_s = 2\sigma_f\,(t/r)$

Cylinder $P_c = \sigma_f\,(t/r)$

where t = initial thickness

r = final radius

Pressure and Time for each forming stage

I.	$P_s = 8$ psi r slowly decreasing	Time = 7 min.
II.	$1/3P_s + 2/3P_c = 27$ psi	Time = 35 min.
III.	$1/3P_s + 2/3P_c = 175$ psi r rapidly decreasing	Time = 23 min.
Post-Forming.	$P = 500$ psi	Time = 12 min.

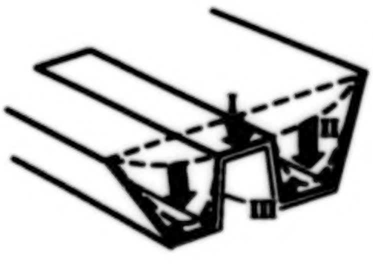


Figure 4. Summary of method used to establish gas pressurization cycle for SPF of 8090 stiffener (plain web configuration) at a constant forming rate.

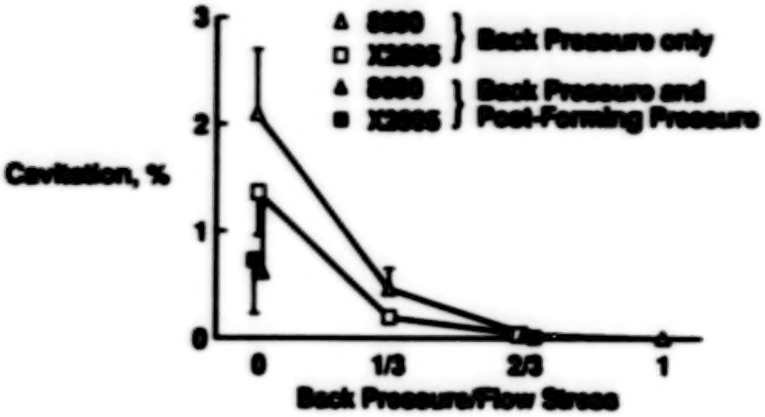


Figure 5. Effect of back pressure and post-forming pressure on cavitation during SPF of 8090 and X2095.

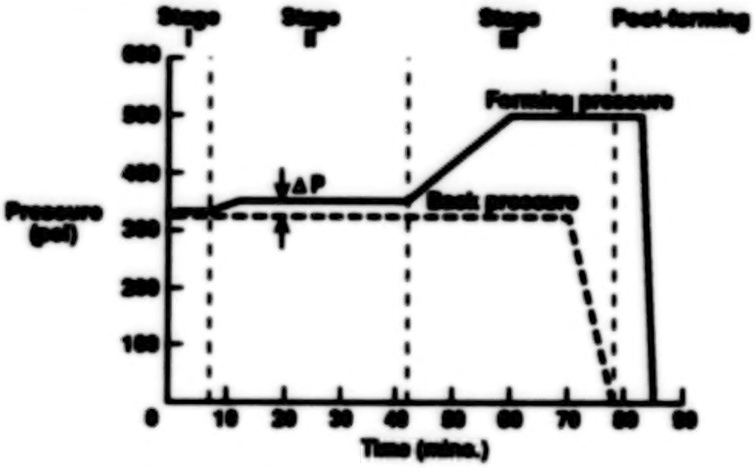


Figure 6. Pressure versus time profile for superplastic forming of 8090 at 985°F corresponding to a strain rate of 2.5×10^{-4} sec^{-1}.

Figure 7. Partially trimmed SPF plain web stiffener components fabricated from 8090, 2090 and X2095 (Weldalite 049). The optimum parameters for forming and post-forming heat treatment are identified for each of the Al-Li alloys.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE IS
OF POOR QUALITY

N91

24066

UNCLAS

N91-24066

LOCALIZED CORROSION OF HIGH PERFORMANCE METAL ALLOYS IN AN ACID/SALT ENVIRONMENT

L. G. MacDowell
NASA Materials Science Laboratory
Kennedy Space Center, FL 32899

C. Ontiveros
California State Polytechnic University
Pomona, CA 91768

ABSTRACT

Various vacuum jacketed cryogenic supply lines at the Space Shuttle launch site at Kennedy Space Center use convoluted flexible expansion joints. The atmosphere at the launch site has a very high salt content, and during a launch, fuel combustion products include hydrochloric acid. This extremely corrosive environment has caused pitting corrosion failure in the thin walled 304L stainless steel flex hoses. A search was done to find a more corrosion resistant replacement material. This study focused on 19 metal alloys. Tests which were performed include electrochemical corrosion testing, accelerated corrosion testing in a salt fog chamber, and long term exposure at a beach corrosion testing site. Based on the results of these tests, several nickel based alloys were found to have very high resistance to this corrosive environment. Also, there was excellent agreement between the electrochemical tests and the actual beach exposure tests. This suggests that electrochemical testing may be useful for narrowing the field of potential candidate alloys before subjecting samples to long term beach exposure.

INTRODUCTION

Flexible hoses are used in various supply lines that service the Space Shuttle Orbiter at the launch pad. These thin walled (.025 in) (.064 cm) convoluted flexible hoses were originally made out of 304L stainless steel. The atmosphere at the launch site has a very high chloride content caused by the proximity of the Atlantic Ocean. During a lunch, the products from the fuel combustion reaction include concentrated hydrochloric acid. This combination of chloride and acid leads to a very corrosive environment. This type of environment causes severe pitting in some of the common stainless steel alloys. In the case of vacuum jacketed cryogenic lines, pinhole leaks caused by failure of the flex hose by pitting produces a loss of vacuum and subsequent loss of insulation. An experimental study was carried out on 19 candidate alloys, including 304L stainless steel for comparison. These alloys were chosen on the basis of their reported resistance to chloride environments. Accelerated corrosion testing and actual field tests were performed with the 19 alloys. This paper summarizes the results of these tests.

EXPERIMENTAL PROCEDURE

Materials

The nominal compositions of the 19 candidate alloys are listed in Table 1. As can be seen from the table, alloys A through I are nickel based alloys. Alloy J is zirconium, and alloys K through S are various stainless steels and other iron based alloys. Alloy K is stainless steel 304L, which is the material originally used to construct the flexible expansion joints.

Electrochemical Tests

A Model 351-2 Corrosion Measurement System, manufactured by EG&G Princeton Applied Research, was used for all electrochemical measurements. Specimens were flat coupons 5/8 in. (1.59 cm) in diameter. The specimen holder is designed such that the exposed metal surface area is 1 cm. The electrochemical cell included a saturated calomel reference electrode (SCE), 2 graphite rod counter electrodes, the metal specimen working electrode, and a bubbler/vent tube. The electrolyte was an aerated solution of HCl plus 3.55wt% NaCl. The concentration of HCl was 0.1N for the first round of testing and was increased to 1.0N for a second round of tests on the more resistant alloys. The solutions were made using deionized water.

Test specimens were polished with 600-grit paper, ultrasonically degreased in a detergent solution, dried, and weighed before immersion in the electrolyte. The electrolyte solution was aerated for at least 45 minutes before immersion of a test specimen. Aeration continued throughout the test. Electrochemical tests performed include determining corrosion potential, polarization resistance, and cyclic polarization. The polarization resistance test procedure was based on ASTM G59. The cyclic polarization procedure was based on ASTM G61. All three electrochemical tests can be run in sequence on a single specimen.

The corrosion potential (Ecorr) was monitored for 3600 seconds, after which time the potential had usually stabilized. For the polarization resistance test, the potential was varied from -20mV to +20mV relative to the measured corrosion potential, while the resulting current was recorded. The scan rate was 0.1 mV/sec. A linear graph of potential vs. current density was made, and the resulting slope (at zero current) plus the Tafel constants were used to calculate the corrosion rate in mils per year (mpy). Tafel constants were calculated using the forward scan of the cyclic polarization data. The cyclic polarization scan started at -250mV relative to Ecorr. The scan rate was 0.166 mV/sec, and the scan was reversed when the current density reached 5 mA/cm^2. The reverse potential scan continued until the potential returned to the starting point of -250mV relative to Ecorr. A graph was then made of potential vs logarithm of current density.

Salt Fog Chamber/Acid Dip

An Atlas Corrosive Fog Exposure System Model SF-2000, manufactured by Atlas Electric Devices Company, was used for accelerated exposure. The solution for salt fog exposure was standard 5% sodium chloride mixture. Specimens were also periodically dipped in a 1.0N hydrochloric acid/alumina mixture. The particles size of the alumina was 0.3 micron. Flat specimens 1in x 2in x 1/8in (2.54cm x 5.08cm x 0.32cm) were used. One set of samples were base metals with an autogenous weld on one end. Another set of samples were the base metal welded to 304L stainless steel. All flat specimens had a 3/8in (0.95cm) hole drilled in the center for mounting purposes. Stress corrosion cracking specimens were also used. These were standard U-bend samples prepared with a weld in the center of the bend.

The flat specimens were weighed on a Mettler AE160 electronic balance. The specimens were then mounted on insulated rods and set in the salt fog chamber at about 15-20 degrees off the vertical. The specimens were exposed to one week of salt fog per ASTM B117. The temperature of the chamber was controlled at 95°F (35°C) . After the one week exposure, the specimens were removed and dipped in the hydrochloric acid/ alumina mixture to simulate the effluent created during launch of the Space Shuttle. After one minute of immersion, the specimens were allowed to drain and dry overnight. The samples were installed in the salt fog chamber for another week, followed by another acid/alumina dip. After a four week/four dip period, the specimens were removed from the mounting rod, cleaned, and weighed. After this inspection, the samples were remounted and returned to the salt fog chamber for another four week/four dip cycle of testing.

Beach Exposure/Acid Spray

All exposure was carried out at the beach corrosion testing site located about 100 feet (30.5m) from the high tide line, along the Atlantic Ocean at Kennedy Space Center, Florida. The metal specimens were a duplicate set as described above for the salt fog/acid dip tests. An acid solution of 10% hydrochloric acid by volume (about 1.0N) mixed with 0.3 micron alumina powder was used. The procedure was based on ASTM G50, with the addition of an acid spray. The specimens were weighed and mounted on short insulated rods that were attached to a plexiglas sheet. The specimens were mounted face side up, facing east towards the ocean at a 45 degree angle and were boldly exposed to the environment to receive the full extent of sun, rain, and sea spray. Approximately every two weeks, the specimens received an acid spray to simulate the effluent during a launch. The acid solution was allowed to remain on the surface of the specimens until it dried or was rinsed off by rain. Periodically, the specimens were removed from the beach, cleaned, and weighed. The samples were then remounted and returned to the beach for continued exposure and acid sprays.

RESULTS

Electrochemical

The electrochemical tests were run first with 3.55 wt% NaCl and 0.1N HCl, measuring only corrosion potential and cyclic polarization data. These tests were repeated, with the insertion of the polarization resistance experiment. There was very good agreement between the two sets of experiments, indicating good reproducibility and a negligible effect of the polarization resistance test on the cyclic polarization results.

Corrosion Potential. Corrosion potential (Ecorr) gives an indication of how noble an alloy is in a given environment. Figure 1 shows the corrosion potential vs. time data for a stable material. Some materials displayed very unstable corrosion potentials, such as shown in Figure 2. Table 2 shows the results for the 19 alloys tested, in order of increasing activity. The potentials are all with respect to the saturated calomel electrode (SCE) reference and were recorded after 1 hour.

Polarization Resistance. Polarization resistance (R_p) is used to calculate the uniform corrosion rate when the potential is close to the corrosion potential. Results of a typical polarization resistance run are shown in Figure 3. The slope of this line is Rp in ohms. Corrosion current density and uniform corrosion rate are then calculated as follows

$$\text{Icorr} = \frac{106\ \text{Ba Bc}}{2.3\ \text{Rp}\ (\text{Ba}+\text{Bc})} \qquad (1)$$

$$\text{Corrosion Rate} = \frac{0.13\ \text{Icorr (E.W.)}}{\text{d}} \qquad (2)$$

where Ba and Bc are the Tafel constants in V/decade, Icorr is the corrosion current density in A/cm^2, E.W. is the equivalent weight in g/equiv, d is the density in g/cm^3, and corrosion rate is mils per year (mpy). Table 3 summarizes the polarization resistance results, with the alloys ranked in order of increasing corrosion rate. The polarization resistance results did not correlate with beach exposure and salt fog chamber results as well as the cyclic polarization results did. In general, the polarization resistance technique works better with metals that display active corrosion behavior in a given environment. It may not give accurate results for passive metal behavior such as many of the alloys displayed during this study. So, for these. exposure conditions, polarization resistance is not the best electrochemical technique to use to predict actual field exposure corrosion results.

Cyclic Polarization. Cyclic polarization gives an indication of a specimen's resistance to pitting corrosion[1,2], and this method as been used for many systems to determine susceptibility to localized corrosion[1-10]. Figure 4 shows a curve with the hysteresis effect typical of a material with a low resistance to pitting. Since the potential scan is at a known constant rate, the potential values can be converted to time, and the area inside the hysteresis loop can be found by integration to give units of $coulomb/CM^2$. This area value should be very small for alloys that are highly resistant to pitting, as seen in Figure 5 which is for a material that is very corrosion resistant. In this case, the reverse scan traces almost exactly over the forward scan. Table 4 ranks the alloys according to area of the hysteresis loop. Visual inspection and inspection under a microscope revealed various levels of pitting corrosion. Crevice corrosion was also observed on several of the samples around the edge of the specimen holder. These visual observations agreed extremely well with the electrochemical results. Some alloys displayed uniform corrosion, rather than localized pitting or crevice corrosion. In these instances, the cyclic polarization curves were similar to the one shown in Figure 6. This type of curve does not yield a meaningful value for hysteresis loop area. Therefore, data for this type of behavior does not appear with the cyclic polarization results.

Increased Acid Concentration. Based on the data in the preceding tables, the most resistant alloys were chosen and run through the same electrochemical tests using a more aggressive electrolyte of 3.55 wt% NaCl with the HCl concentration increased to 1.0N. Stainless steel 304L was also used, as a basis for comparison. Table 5 shows the effect on corrosion potential of increasing the acid concentration. Table 6 shows the polarization resistance results obtained with the higher acid concentration. As with the results in Table 3, this test separates out some of the poor performers, but it can not rank the alloys accurately. The cyclic polarization results for the stronger electrolyte are summarized in Table 7, with the alloys ranked according to weight loss. The 304L sample experienced uniform corrosion. Therefore, results for 304L do not appear in Table 7, for the reason mentioned in regard to Figure 6 (i.e., no meaningful value for hysteresis loop area). Alloys R and Q suffered severe uniform corrosion, in addition to pitting, which is why the area values for these two alloys do not correlate with the weight loss values. Since there was uniform corrosion, the area values are not really meaningful, and the weight loss gives a better indication of the extent of corrosion for these two alloys.

Salt Fog Chamber/Acid Dip

Corrosion rates were calculated by

$$\text{Corrosion Rate (mpy)} = \frac{534\,w}{dAt} \qquad (3)$$

where w is the weight loss in mg, d is the metal density in g/CM^3, A is the area of exposure in in^2, and t is the exposure time in hours. This expression calculates the uniform corrosion rate over the entire surface and gives no indication of the severity of localized attach (pitting and/or crevice corrosion). Specimens were also examined visually for signs of localized corrosion. Table 8 summarizes the weight loss and corrosion rate results after 8 weeks and after 20 weeks. These results are for the autogenous weld samples.

In conjunction with the standard alloy coupons, specimens welded to 304L stainless steel were also tested. This was done since any new replacement alloy would be installed in an existing 304L stainless steel piping system, and galvanic corrosion in the weld area could become a source of system failure. Most of these specimens suffered some type of weld decay. In general, the deterioration was mainly on the 304L surfaces adjacent to the weld. Since the particular application of a new corrosion resistant alloy would be to form thin wall convolutes welded to a heavy wall 304L stainless steel pipe, the galvanic effect should be minimal.

Beach Exposure/Acid Spray

Corrosion rates were calculated using equation 3. The weight loss and corrosion rate data are shown in Table 9 for 60 days, 251 days, and 479 days of exposure. As can be seen from the table, several materials clearly separated from the rest and displayed excellent corrosion resistance. Alloys B, F, C, and A showed virtually no weight loss after more than 15 months on the beach and 30 acid sprays. These are all nickel based alloys. The various stainless steel alloys all showed considerably higher corrosion rates.

CONCLUSIONS

There was excellent agreement between the cyclic polarization results and the long term beach exposure results. With the exception of the zirconium sample (alloy J), which performed quite well at the beach but not under cyclic polarization conditions, the top six alloys from the cyclic polarization tests are the same as the beach exposure results. This supports the claim that cyclic polarization is a good way to determine an alloy's resistance to localized corrosion in a given electrolyte. Cyclic polarization is very quick compared to long term beach exposure. This accelerated test can be used to screen prospective alloys before exposing them to actual beach conditions.

When the beach results are compared to the salt fog results, many materials change positions relative to each other. In general, though, the materials at the top and at the bottom of each list remained in their respective positions. Beach testing should be considered the best judge of alloy's performance since it has naturally occurring conditions. However, the accelerated testing does give insight into which metals have a good chance of performing well.

Finally, several alloys were found that have superior resistance to pitting and crevice corrosion in an acidic salt environment. Alloy B was chosen and has been used to replace some of the thin walled flexible expansion joints at the Space Shuttle launch pad.

REFERENCES

1. ASTM G61-78, Standard Practice for Conducting Cyclic Potentiodynamic Polarization Measurements For Localized Corrosion, 1986 Annual Book of ASTM Standards, Volume 03.02, ASTM, Philadelphia, PA, 1986.

2. R. Baboian, G.S. Haynes, Cyclic Polarization measurements - Experimental Procedure and Evaluation of Test Data, in Electrochemical Corrosion Testing, ASTM STP 727, F. Mansfeld, U. Bertocci, Eds., ASTM, Philadelphia, PA, p. 274, 1981.

3. W. F. Bogaerts, A.A. Van Haute, Corrosion Science, Vol., 25, No. 12, p. 1149, 1985.

4. A. Kawashima, K. Hashimoto, Corrosion Science, Vol 26, No. 6, p. 467, 1986.

5. B.E. Wilde, Corrosion, Vol. 42, No. 3, p. 147, 1986.

6. P.B. Lindsay, Materials Performance, Vol. 25, No. 12, p. 23, 1986.

7. A.I. Asphahani, Materials Performance, Vol. 19, No. 8, P. 9, 1980.

8. R. Bandy, D. Van Rooyen, Corrosion, Vol. 39, No. 6, p. 227, 1983.

9. P.E. Manning, Corrosion, Vol. 39, No. 3, p. 98, 1983.

10. E.L. Liening, Materials Performance, Vol. 19, No. 2, p. 35, 1980

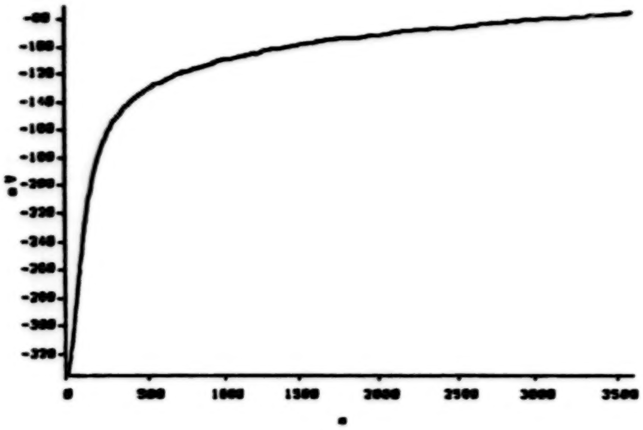


Figure 1 Stable Corrosion Potential
Alloy R in Aerated 3.55%NaCl + 0.1NHCl
Corrosion Potential (mV) vs Time (sec)

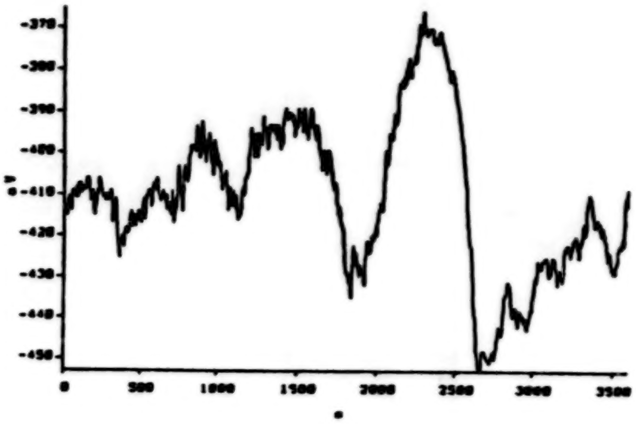


Figure 2 Unstable Corrosion Potential
Alloy L in Aerated 3.55%NaCl + 0.1NHCl
Corrosion Potential (mV) vs Time (sec)

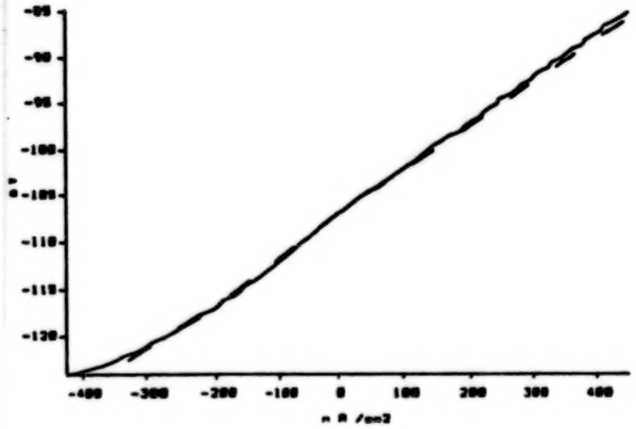


Figure 3 Polarization Resistance Graph
Alloy A in Aerated 3.55%NaCl + 0.1NHCl
Potential(mV) vs Current Density(nA/cm^2)

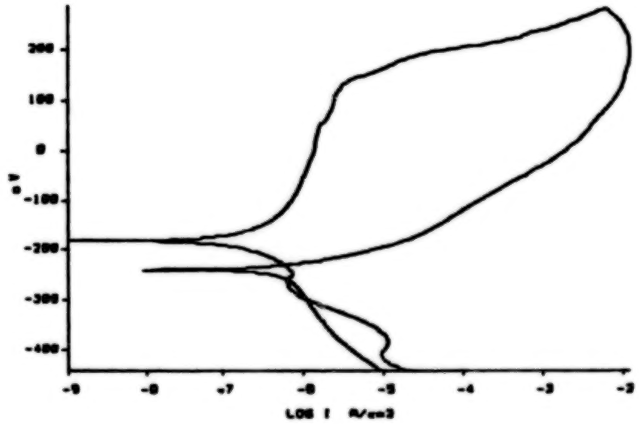


Figure 4 Cyclic Polarization With Hysteresis Loop
Alloy M in Aerated 3.55%NaCl + 0.1NHCl
Potential (mV) vs log I(A/cm^2)

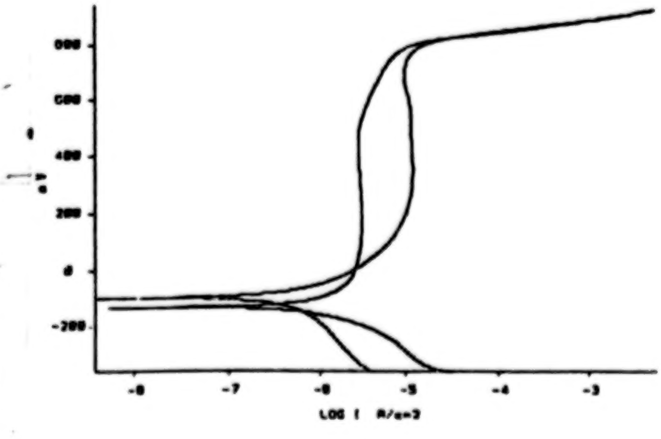


Figure 5 Cyclic Polarization Without Hysteresis Loop
Alloy A in Aerated 3.55%NaCl + 0.1N HCl
Potential (mV) vs log I(A/cm^2)

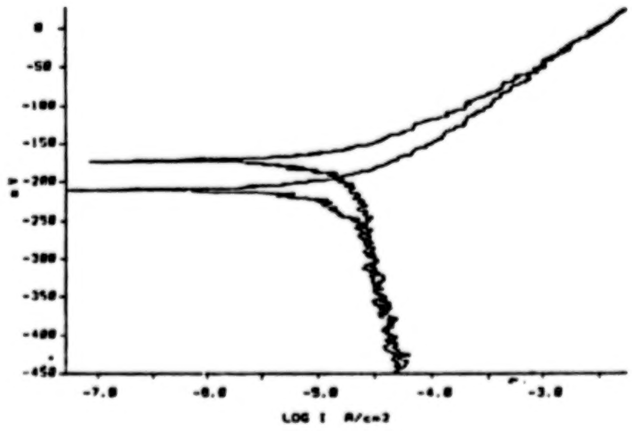


Figure 6 Cyclic Polarization With Uniform Corrosion
Alloy I in Aerated 3.55%NaCl + 0.1N HCl
Potential (mV) vs Log Current Density (A/cm^2)

ORIGINAL PAGE IS
OF POOR QUALITY

TABLE 1
NOMINAL COMPOSITIONS (WT%) OF THE CANDIDATE ALLOYS

Alloy	Ni	Fe	Cr	Mo	Mn*	Cu	C*	Si*	Other
A	Bal.	3.0	18	17	1.0		.01	.08	
B	Bal.	3.0	22	13	0.5		.01	.08	V 0.3,W 3
C	Bal.	7.0	17	17	1.0		.01	.08	V 0.3,W 4.5
D	Bal.	2.0	1	28	1.0		.01	.10	
E	Bal.	8.0	16		1.0	0.5	.15	.50	
F	Bal.	5.0	23	10	0.5		.10	.50	
G	Bal.	22	21	3	1.0	2.5	.05	.50	
H	Bal.	20	22	7	1.0	2.0	.02	1.0	W 1.5
I	Bal.	2.5			2.0	31	.30	.50	
J									Zr 99.2
K (304L)	10	Bal.	19		2.0		.03	1.0	
L (304LN)	10	Bal.	19		2.0		.03	1.0	
M (316L)	12	Bal.	17	2.5	2.0		.03	1.0	
N (317L)	13	Bal.	19	3.5	2.0		.03	1.0	
O (904L)	25	Bal.	21	4.5	2.0	1.5	.02	1.0	
P	35	Bal.	20	2.5	2.0	3.5	.07	1.0	
Q	4	Bal.	28	2.0	2.0		.03	.60	
R	5	Bal.	22	3.0	2.0		.03	1.0	
S	5	Bal.	26	3.0	1.5	2.0	.04	1.0	

* Maximum

TABLE 2
CORROSION POTENTIAL IN 3.55% NaCl + 0.1N HCl

Alloy	Ecorr(mV)	Alloy	Ecorr(mV)
P	- 57	N	-121
Q	- 72	F	-145
B	- 77	D	-159
R	- 83	M	-170
H	-100	I	-175
O	-100	E	-272
A	-105	J	-319
C	-109	K	-407
G	-113	L	-414
S	-120		

All potentials are with respect to the saturated calomel reference electrode and are the average of 2 or more runs.

TABLE 3
POLARIZATION RESISTANCE RESULTS IN 3.55% NaCl + 0.1N HCl

Alloy	Rp(ohms)	Icorr(amps)	Corrosion Rate (mpy)
G	180000	1.15E-7	0.05
S	247000	1.23E-7	0.06
Q	240000	1.52E-7	0.07
J	248000	1.89E-7	0.09
H	140000	2.61E-7	0.12
O	150000	2.57E-7	0.12
P	98500	3.08E-7	0.14
R	127000	3.31E-7	0.15
N	131000	3.46E-7	0.16
M	125000	3.47E-7	0.16
B	84000	5.06E-7	0.22
F	90000	5.42E-7	0.24
A	47500	1.11E-6	0.47
C	44000	1.45E-6	0.62
L	6450	3.70E-6	1.69
D	1800	1.01E-5	4.16
I	930	3.00E-5	13.00
E	665	3.68E-5	16.50
K	352	5.80E-5	26.00

TABLE 4
AREA OF HYSTERESIS LOOP WITH 3.55% NaCl + 0.1N HCl

Alloy	Area(Coulombs)	Alloy	Area(Coulombs)
B	2.0	O	7.0
C	2.0	J	9.0
A	2.0	G	9.0
H	2.0	K	12.0
S	3.0	L	14.0
F	3.0	M	15.0
E	4.0	N	17.0
Q	4.0	P	22.0
R	5.0		

TABLE 5
CORROSION POTENTIAL IN 1.0N HCl AND IN 0.1N HCl
(BOTH WITH 3.55% NaCl)

Alloy	Ecorr(mV) 1.0N	Ecorr(mV) 0.1N
H	- 51	-100
P	- 58	- 57
A	- 98	-105
B	-106	- 77
C	-108	-109
O	-328	-100
S	-421	-120
R	-422	- 84
K	-452	-408
Q	-455	- 72

All potentials with respect to the SCE reference

TABLE 6
POLARIZATION RESISTANCE RESULTS IN 3.55% NaCl + 1.0N HCl

Alloy	Rp(ohms)	Icorr(amps)	Corrosion Rate (mpy)
S	43000	2.00E-7	0.09
P	84400	4.62E-7	0.21
H	77500	5.31E-7	0.24
B	61700	7.21E-7	0.31
C	38700	1.32E-6	0.56
A	23800	2.34E-6	1.00
O	257	7.88E-5	36.80
K	167	9.56E-5	43.70
R	79	3.85E-4	180.00
Q	37	9.82E-4	457.00

TABLE 7
CYCLIC POLARIZATION RESULTS IN 3.55% NaCl + 1.0N HCl

Alloy	Area(Coulombs)	Wt Loss(mg)
H	1.0	0.2
B	1.0	0.3
C	1.0	0.3
A	1.0	0.3
S	2.0	0.4
F	3.0	0.6
O	7.0	1.6
R	2.0	2.8
Q	1.0	6.9

TABLE 8
RESULTS OF EXPOSURE TO 5% SALT FOG PLUS ACID DIPS

	8 weeks + 8 Acid Dips		20 weeks + 20 Acid Dips	
Alloy	Wt Loss(mg)	Corr. Rate(mpy)	Wt Loss(mg)	Corr. Rate(mpy)
B	1.5	.015	0.9	.004
F	2.7	.027	2.5	.010
J	1.2	.016	2.0	.011
C	2.8	.026	3.5	.013
A	2.9	.028	3.7	.014
H	7.1	.073	9.3	.038
D	42.0	.382	154.7	.563
S	93.9	1.045	158.1	.704
O	69.5	.728	179.5	.753
G	85.4	.893	185.8	.778
L	62.0	.605	228.8	.893
R	128.6	1.150	251.8	.900
N	69.9	.752	212.2	.909
K	67.2	.690	226.9	.932
Q	91.6	1.035	207.2	.937
E	91.5	.942	229.8	.947
M	63.1	.673	227.6	.971
P	170.5	1.830	374.6	1.611
I	190.8	1.875	619.6	2.435

TABLE 9
RESULTS OF EXPOSURE TO BEACH CORROSION SITE PLUS ACID SPRAYS

	60 Days + 5 Sprays		251 Days + 13 Sprays		479 Days + 30 Sprays	
Alloy	Wt. Loss (mg)	Corr. Rate (mpy)	Wt. Loss (mg)	Corr. Rate (mpy)	Wt. Loss (mg)	Corr. Rate (mpy)
B	0.0	.000	0.0	.000	0.0	.000
F	0.0	.000	0.0	.000	0.3	.000
C	0.1	.001	0.1	.001	0.5	.001
A	0.1	.001	0.1	.001	0.7	.001
H	1.5	.014	3.4	.008	3.7	.004
J	0.7	.008	1.4	.004	8.1	.012
S	10.5	.110	13.9	.034	16.3	.021
R	12.1	.099	25.1	.049	37.1	.038
Q	13.0	.139	22.0	.056	35.9	.048
O	14.7	.144	29.3	.069	45.3	.056
G	12.4	.120	28.8	.068	55.5	.069
E	20.3	.195	49.7	.114	81.0	.097
N	18.8	.187	45.0	.107	78.6	.097
M	24.7	.245	56.6	.134	106.2	.132
K	27.7	.278	61.2	.147	113.5	.143
L	34.8	.320	81.6	.177	145.9	.166
D	32.9	.280	106.4	.218	205.1	.220
P	43.1	.435	107.4	.259	222.9	.282
I	95.4	.871	244.7	.534	481.0	.550

SESSION F - OPTICS AND COMMUNICATIONS

Wednesday November 28, 1990

- **Digital Codec For Real-Time Processing Of Broadcast-Quality Video Signals**
- **Recent Advances In Coding Theory For Near-Error-Free Communications**
- **Microwave Integrated Circuits For Space Applications**
- **Optical Communications For Space Missions**
- **Optical Shutter Switching Matrix**
- **Fiber Optic Tactical Local Area Network (FOTLAN)**
- **High-Precision Applications Of The Global Positioning System**

PRECEDING PAGE BLANK NOT FILMED

N91

24067

UNCLAS

N91-24067

Digital CODEC for Real-time Processing of Broadcast Quality Video Signals at 1.8 Bits/Pixel

Mary Jo Shalkhauser, Wayne A. Whyte
NASA Lewis Research Center
21000 Brookpark Road
Mail Stop 54-2
Cleveland, Ohio 44135
TEL: 216-433-3455
TEL: 216-433-3482
FAX: 216-433-8705

Abstract

Advances in very large-scale integration and recent work in the field of bandwidth efficient digital modulation techniques have combined to make digital video processing technically feasible and potentially cost competitive for broadcast quality television transmission. A hardware implementation has been developed for a DPCM-based digital television bandwidth compression algorithm which processes standard NTSC composite color television signals and produces broadcast quality video in real time at an average of 1.8 bits/pixel. This paper describes the data compression algorithm and the hardware implementation of the codec, and provides performance results.

Introduction

Transmission of television signals in a digital format has been looked upon with promise for a number of years. Digital systems providing teleconferencing quality video have become common place in both government and industry. However, digital transmission of high-quality (toll grade or broadcast quality) television signals has yet to achieve anything close to the same kind of acceptance. This has been due in part to the broadcasters' reluctance to have processing of any kind performed on the transmitted signals. But to a greater extent, digital transmission of broadcast quality video has failed to gain acceptance because it has not been cost effective to do so. The lack of available wideband digital links as well as the complexity of implementation of bandwidth efficient digital video CODECs (encoder/decoder) has worked to keep the cost of digital television transmission too high to compete with analog methods.

Advances in very large-scale integration (VLSI) as well as recent work in the field of advanced digital modulation techniques have combined to make digital video processing technically feasible and potentially cost competitive for broadcast quality television transmission. The coupling of a transparent, bandwidth efficient data compression technique with a bandwidth efficient modulation technique offer the potential for transmission of two (or more) high-quality television signals in the same bandwidth occupied by a single frequency modulated television signal. This paper presents the hardware implementation of a digital television bandwidth compression algorithm which processes standard NTSC (National Television Systems Committee) composite color television signals and produces broadcast quality video in real time at an average of 1.8 bits/pixel. (A pixel, or picture element, represents each piece of sampled data. The sampling rate used with this algorithm results in 768 samples over the active portion of each video line by 512 active video lines per video frame.) The algorithm is based on differential pulse code modulation (DPCM), but additionally utilizes a non-adaptive predictor, non-uniform quantizer and multilevel Huffman coder to reduce the data rate substantially below that achievable with straight DPCM. The non-adaptive predictor and multilevel Huffman coder combine to set this technique apart from prior-art DPCM encoding algorithms. The following sections will provide the details of the compression algorithm, the hardware implementation and performance results, respectively.

The relative simplicity of the encoder and decoder hardware, direct interfacing to NTSC composite video signals, and exceptional performance make the system attractive to a wide variety of applications. The CODEC may be used in virtually any application where high quality video transmission is required; conventional television broadcasting, cable distribution to subscribers, video trunking for distribution via commercial communications satellites, and direct-to-the-user satellite broadcasting (DBS).

Data Compression Algorithm

Differential pulse code modulation has historically been one of the most popular predictive image coding methods studied, due to its simplicity of implementation and overall subjective performance characteristics. The fault of DPCM schemes in the past have been that 3-4 bits/pixel were required to achieve acceptable image quality, with 4 bits/pixel generally preferred to maintain a broadcast quality picture representation. The system presented here combines the simplicity of the basic DPCM approach with several performance enhancements to achieve broadcast quality images at an average 1.8 bits per pixel.

A block diagram of the compression scheme is presented in figure 1. The DPCM portion utilizes an intrafield approach with a two dimensional prediction based on averaging neighboring pixel values having the same color subcarrier phase relationship as the current pixel. Sampling of the composite analog video signal is done at four times the color subcarrier frequency rate (4 * 3.579545 MHz). Two pixels are averaged to generate the predicted value, PV, in figure 1; the fourth previous pixel from the same line and the same pixel from two lines previous in the same field. These neighboring pixels have the same color subcarrier phasing as the current pixel and will therefore be highly correlated. The two pixel values are averaged to produce the prediction of the current pixel value. At this point the algorithm differs from standard DPCM, where the predicted value would simply be subtracted from the current pixel value to obtain a difference value to be quantized.

Figure 1 shows a "non-adaptive predictor" value (NAP) being subtracted from the current pixel value along with the predicted value, PV. The function of the NAP is to further improve the prediction of the current pixel. The non-adaptive predictor estimates the difference value obtained when the DPCM prediction is subtracted from the current pixel value (PIX - PV). The subtraction of the NAP value from PIX - PV causes the resulting difference value (DIF) to be close to zero. The smaller the DIF, the more efficiently the quantized pixel information can be transmitted due to the use of Huffman coding prior to transmission over the channel. (Huffman coding assigns variable length codewords based upon probability of occurrence. The application of Huffman coding to this algorithm will be discussed later.)

The development of the non-adaptive predictor was predicated on the likelihood that the difference values of adjacent pixels are similar. The difference between the current pixel value and its prediction, PV, is estimated and subtracted off by way of the NAP prior to quantization. The estimate is simply based on the value of DIF for the previous pixel. The NAP is non-adaptive because the estimates are prestored and do not change with differing picture content. These prestored values were generated from statistics of numerous television images covering a wide range of picture content. The NAP values represent the average difference values calculated within the boundaries of the difference values for each quantization level. Table 1 shows the NAP values corresponding to each quantization level. To give an example using the values in Table 1; if the difference value (DIF) for the previous pixel was 40, corresponding to quantization level 11, the value of NAP to be subtracted off from the current pixel difference would be 38. To reconstruct the pixel, the decoder uses a lookup table to add back in the appropriate NAP value based upon knowledge of the quantization level from the previously decoded pixel. The use of the NAP results in faster convergence at transition points in the image, thereby improving edge detection performance. The rapid convergence also reduces the total data requirements by increasing the percentage of pixels in quantization level 7, which is assigned the shortest codeword length by the Huffman coding process.

The quantizer shown in figure 1 has thirteen (13) levels. Each level has a quantization value associated with a range of difference values as indicated in Table 1. The quantizer is non-uniform so that more levels are provided for small magnitude differences which would result from subtle changes in picture content. The human eye is sensitive to small variations in smooth regions of an image and can tolerate larger variations near transition boundaries where large difference values are more likely to occur. The non-adaptive predictor discussed previously, acts to reduce the difference values thus improving image quality by reducing the quantization error. This is because the non-uniform quantizer results in lower quantization error for small magnitude differences than for large magnitude differences. The number of quantization levels, the corresponding difference value ranges, and the specific quantization values shown in Table 1 were experimentally derived through subjective evaluation of sample images processed by computer simulation of the encoding algorithm.

The final major aspect of the encoding algorithm is the multilevel Huffman coding process. Huffman coding of the quantized data allows shorter codewords to be assigned to quantized pixels having the highest probability of occurrence. A separate set of Huffman codes has been generated for each of the thirteen quantization levels. The matrix of code sets is used to reduce the number of data bits required to transmit a given pixel. The particular Huffman code set used for a given quantized pixel is determined by the quantization level of the previous pixel (i.e. if the difference value for the previous pixel resulted in quantization level 4 being selected for that pixel, then the Huffman code set selected for the current pixel would be code set 4, corresponding to the probability of occurrence of pixels falling into the fourth quantization level).

As with the NAP, the Huffman code trees were generated by compiling statistical data from numerous images covering a broad range of picture content. Probability of occurrence data was compiled for each of the thirteen quantization levels as a function of the quantization level of the previous pixel. A separate Huffman code set was then generated based on the probability data of "current" pixels falling into each of the thirteen quantization levels of the "previous" pixels. There is a tendency for neighboring pixels to fall into the same or close to the same quantization level. By recognizing and taking advantage of this fact, the use of the multilevel Huffman code sets provides significant reductions in bits per pixel over a single Huffman code tree because they allow nearly all pixels to be represented by short length codewords.

Hardware Implementation

The configuration of the data compression hardware is shown in figure 1. The NTSC format video signal is digitized by sampling with an analog-to-digital (A/D) converter at a rate of four times the color subcarrier frequency (approximately 14.32 MHz). The A/D converter has an 8-bit resolution allowing each sample, called a pixel, to be represented by one of 256 digital levels. The eight-bit pixels are input to the encoder at the 14.32 MHz rate. The encoder compresses the video data and serially transmits a compressed representation of the data over a channel at a rate of approximately 25 megabits per second (Mbps) to the decoder.

The decoder receives the compressed serial data and reconstructs a facsimile of the original video data. The reconstructed 8-bit pixels are converted back to an analog video signal using a digital-to-analog (D/A) converter. For this implementation, the entire video signal including the horizontal and vertical synchronization pulses, the color burst, and the active video is sampled and compressed. In future implementations, all but the active video part of the signal will be removed by the encoder and the decoder will reconstruct these portions and reinsert them into the signal, thereby increasing the overall compression of the picture.

For this breadboard version of the compression hardware, TTL (transistor-transistor logic) was chosen as the implementation technology due to ease of usage and wide availability of devices. The hardware designs were constructed on wire-wrap boards and mounted in a five-slot 19-inch chassis.

ENCODER

The encoder portion of the compression hardware digitizes the analog video signal into 8-bit pixels, compresses the image, and converts the resultant data to a serial bit stream. A detailed description of each of the encoder blocks (shown in figure 1) follows.

The differential pulse code modulation (DPCM) circuit averages previous neighboring pixel values to predict the current pixel value. The previous pixels of the same color subcarrier phase as the current pixel are obtained using a 4-pixel delay and a two-line delay. The 4-pixel delay is implemented using four 8-bit registers in a shift register configuration.

The 2-line delay is implemented using a random access memory (RAM) which is addressed by a counter that recycles every two lines. For the first two lines of each field, the RAM is loaded with the reconstructed values of the original pixels, while the output register of the 2-line delay is zeroed. For every line thereafter, the pixel value of two lines previous is read out of the RAM and then the new reconstructed pixel value (RP) is written into the same memory location. Then the address counter is incremented to the next memory location for the next pixel prediction.

The outputs of the 2-line delay and the 4-pixel delay are added together using two cascaded 4-bit full adders. The divide-by-two function is performed by dropping the adder's least significant output bit and using the carry-out signal as the most significant bit. This is the same as a "shift right with carry" of the adder outputs. During the first two lines of each field, the DPCM prediction circuit uses only the 4th previous pixel on the current line to predict the current pixel value. In this case, the 2-line delay input to the adder is zeroed and the divide-by-two circuit is by-passed using a multiplexer.

The DPCM predicted value (PV) and the non-adaptive prediction value (NAP) are subtracted (subtractions are performed as two's complement additions) from the original pixel value resulting in a difference value (DIF = PIX-PV-NAP). These difference values are then grouped into quantization levels (QL), created from a look-up table implemented in programmable read-only memory (PROM) using the difference value (DIF) as the address. The quantization levels are delayed by one pixel-time and used to address another PROM look-up table to create the non-adaptive prediction (NAP). The non-adaptive predictor estimates the current DPCM difference value (PIX-PV) from the difference value of the immediately previous pixel.

The quantization value (QV), an estimation of the difference value (DIF), is created from another PROM look-up table. In this case, the quantization level is used to address the memory locations which contain the quantization values.

The two dimensional Huffman codes are created by yet another PROM look-up table. The current quantization level (QL) and the immediately previous quantization level (QL_{n-1}) address a PROM which contains, at each location, a 1 to 12-bit Huffman code and a 4-bit code which specify the length of the Huffman code.

The outputs of the Huffman encoder are multiplexed with the first four pixels of every line so that the DPCM circuit has a valid starting point. The output of this multiplexer is input into a bank of first-in, first-out (FIFO) memories. Forty FIFO integrated circuits are configured with expanded width and depth to achieve a bank of FIFO memory 18-bits wide and 72K deep. The FIFOs are necessary to compensate for the variable lengths of the Huffman codes and the differences between the FIFO input frequency and the FIFO output frequency. On the input-side of the FIFOs, the data is written periodically at the pixel rate of 14.32 MHz. On the output side of the FIFOs, data is read out at a variable rate depending on the length of the Huffman codes and the frequency of the serial data.

Sixteen of the FIFO's bits are data (either actual pixel values for the first four pixels of each line or Huffman codes) and length of data. The other two bits are used to pass line and field flags, indicating the start of each line and each field. The line and field flags are used for insertion of unique words into the data.

Unique words are necessary to maintain proper field and line timing at the decoder. Due the variable length nature of the Huffman codes, channel bit errors can often result in improper detection of the codes by the decoder. Unique words allow the line and field timing to get back on track in the event of bit errors, to minimize the impact to the quality of the reconstructed video images. Different unique word values are used for lines and fields so they can be detected separately. In both cases, unique words were chosen to avoid duplication by valid Huffman codes. Sixteen-bit unique words are currently used; however, the hardware was flexibly designed so that the unique word content and length can be changed if necessary.

The line and field flags at the FIFO outputs are monitored to allow insertion of the unique words at the proper position within the data. When a line or field flag is detected, FIFO reads are stopped to allow time for the unique words to be multiplexed within the data. Like the Huffman codes, the unique words must contain a 4-bit code indicating the length of the unique words. The unique words are divided into two 8-bit sections each accompanied by a length code. After insertion of the unique word, the FIFO reads are reactivated.

Next, the data must be converted from the parallel format to a serial format for transmission over a channel. Because lengths of the Huffman codes vary, a variable length parallel-to-serial converter must be used. The parallel-to-serial converter consists of a 12-bit parallel load shift register and a counter. The Huffman codes are loaded into the shift register and the 4-bit length of the Huffman code is loaded into the counter. The counter counts down as the shift register shifts out the data into a serial bit stream. When the counter reaches zero the shifts stop and a new code is read from the FIFO memory. Then the shift register and counter are loaded with new values and the shifting process repeats.

DECODER

The decoder circuit receives the serial data that the encoder transmitted and reconstructs a representation of the original 8-bit pixels, and using a digital-to-analog (D/A) converter, creates an analog video signal. A detailed description of each of the decoder blocks (figure 1) follows.

The input to the decoder contains three parallel circuits: line unique word detect, field unique word detect, and Huffman decoder. The unique word detect circuits allow detection of unique words with bit errors by selection of an error threshold of up to 3 bit errors. The serial data is first shifted into a 16-bit shift register. The 16-bit parallel output of the shift register is exclusive-or'd (XOR) with the correct unique word value set in dip switches. Next, the number of ones contained in the XOR outputs are summed using adders. A circuit at the output of the adders allows selection of the error threshold and creates a pulse if a unique word within that error threshold is detected. The unique word detect pulse is ANDed with a unique word window signal which disallows unique word detects until close to the expected location of valid unique words. The windowing technique lowers the probability of false detects.

The Huffman decoder receives the data in a serial format from the output of a 16-bit shift register that parallels the unique word detect circuits. When unique words are detected, the Huffman decoder is disabled while the unique word is purged from the shift register to avoid Huffman decoding of unique words.

The Huffman decoder is implemented as a tree search in memory. The address to the Huffman decoder PROM is initially set to zero (top node of the Huffman code tree). The content of each memory location consists of the next two possible addresses to the memory denoting the next two tree branches. As each serial bit is received, it is used by a multiplexer to select the next memory address. A serial "one" selects one address (branch) and a serial "zero" selects the other address (branch). The new address (new tree node) also contains the next two possible tree branches based upon the next received serial bit. The tree search continues in this manner until the least significant output bit of the memory is high, indicating the end of a valid Huffman code. At this point, the other memory output bits contain the correct quantization value (QV) and quantization level (QL) for the received Huffman code. The PROM address is then reset to zero (the top node of the tree) and the decoding process continues.

As the Huffman codes are detected, the resultant quantization levels and values are written into FIFOs. These FIFOs, like in the encoder, are required to absorb the differences in the variable length Huffman codes and the pixel rate at the output of the decoder circuit. In conjunction with the unique word detects and the line and field counters, the FIFO writes and reads are controlled to compensate for synchronization problems created by improper Huffman decoding due to bit errors.

The FIFO outputs, quantization levels and quantization values, are used to reconstruct the original video image. The quantization level is delayed by one pixel-time and used by a PROM look-up table to create the non-adaptive prediction (NAP). The quantization value (QV) is added to the non-adaptive prediction value (NAP) and the DPCM prediction value (PV) to create the reconstructed pixel values (RP). The decoder DPCM circuit implementation is identical to the encoder DPCM circuit. The RP values are input to a D/A converter which converts the reconstructed pixel values to an analog video signal.

Performance Results

Table 2 provides results of computer simulation of the encoding algorithm described in this paper. The picture content across the images is representative of the broad range of material which makes up typical television viewing. The results show the variability of compression with complexity of picture content. Standard color bars, containing considerable redundancy, can be processed very efficiently at 1.347 bits per pixel (bpp). The Beach scene, one of the Society of Motion Picture and Television Engineers (SMPTE) color reference subjective testing slides, requires 2.228 bpp; an indication of the complex nature of the scene. The other images fall somewhere between these bounds, with an average across all scenes of 1.822 bpp. Figure 2 shows the original and reconstructed images.

The reconstructed image quality in all cases is excellent - reconstructed images are indistinguishable from the 8 bpp digitized originals. A two channel video frame storage unit was used in the comparison of original versus processed images. This provided a means for very critical comparative testing, since side-by-side as well as switched comparisons were possible. Interframe motion is not degraded by the processing since the algorithm is an intrafield coding process. Motion sequences of "real-time" video were processed on a frame-by-frame basis and pieced together using a type "C" one-inch broadcast video tape recorder with animation editing capabilities. No motion artifacts were present in the reconstructed sequence as expected.

The hardware design of this video data compression algorithm is relatively straight forward and can be easily implemented in VLSI for performance improvement, size reduction and cost effectiveness. Combining the data compression hardware with a bandwidth efficient modulation technique such as 8-PSK (practical bandwidth efficiency of greater than 2 bits/sec/Hz in a power limited system) will enable two or more broadcast quality television signals to be transmitted through a single C-Band transponder, creating a substantial bandwidth improvement over analog television transmissions.

The relative simplicity of the encoder and decoder hardware, direct interfacing to NTSC composite video signals, and exceptional performance make the system potentially attractive to a wide variety of applications. The CODEC may be used in virtually any application where high quality video transmission is required; conventional television broadcasting, cable distribution to subscribers, video trunking for distribution via commercial communications satellites, and direct-to-the-user satellite broadcasting (DBS). The algorithm embodied by the CODEC is also applicable to bandwidth compression of high definition television signals.

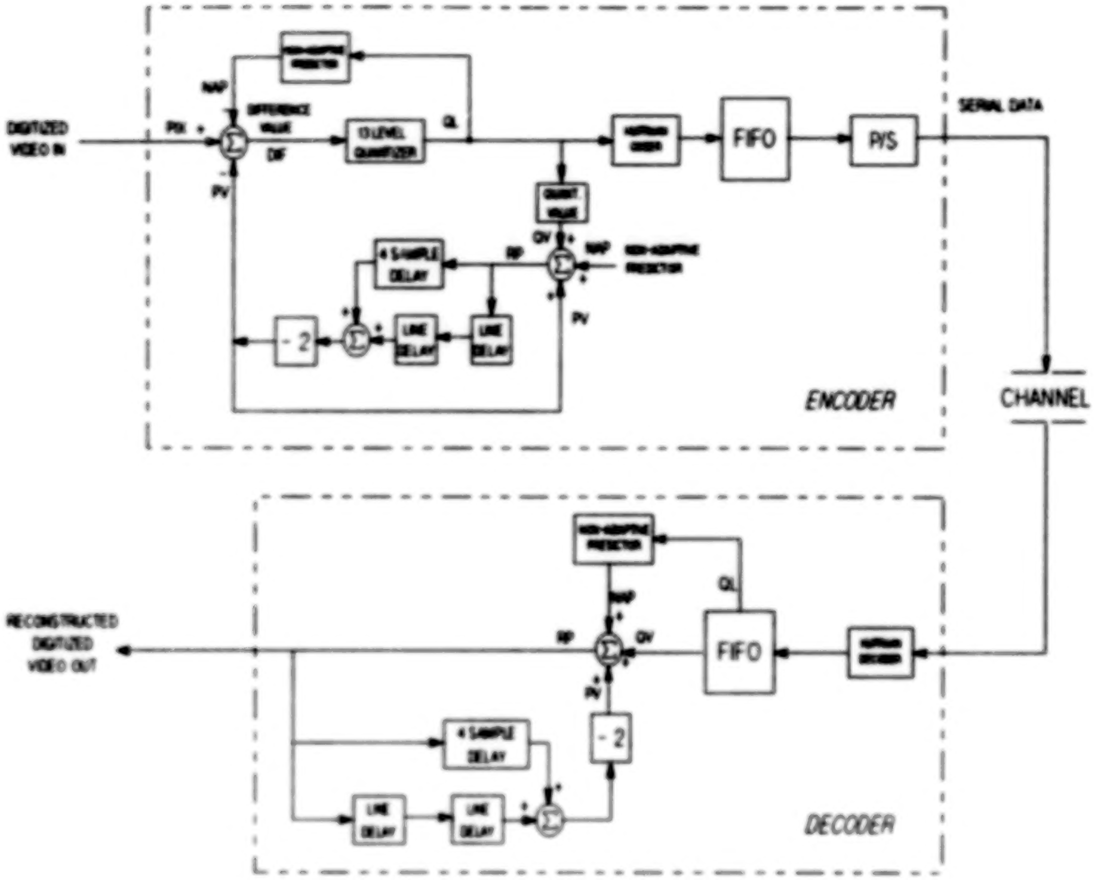

DIGITIZED VIDEO IN
PIX
DIF
QL
NAP
PV
FIFO
P/S
SERIAL DATA
QV
RP
4 SAMPLE DELAY
LINE DELAY
LINE DELAY
- 2
ENCODER
CHANNEL
RECONSTRUCTED DIGITIZED VIDEO OUT
QL
NAP
RP
QV
PV
FIFO
4 SAMPLE DELAY
LINE DELAY
LINE DELAY
- 2
DECODER

Figure 1 - Compression Algorithm Block Diagram

Table 1 - Quantization and Non-adaptive Prediction.

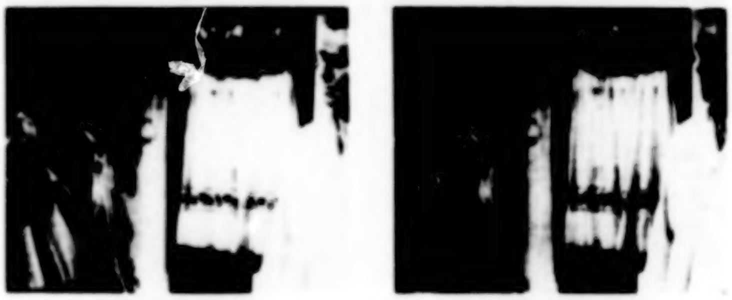
DIFFERENCE VALUE	QUANTIZATION LEVEL	QUANTIZATION VALUE	NON-ADAPTIVE PREDICTION
-255 TO -86	1	-100	-85
-85 TO -60	2	-66	-61
-59 TO -34	3	-42	-38
-33 TO -19	4	-25	-22
-18 TO -9	5	-14	-11
-8 TO -4	6	-6	-4
-3 TO 3	7	0	0
4 TO 8	8	6	4
9 TO 18	9	14	11
19 TO 33	10	25	21
34 TO 59	11	42	38
60 TO 85	12	66	61
86 TO 255	13	100	84

Table 2 - Encoding Algorithm Performance Results

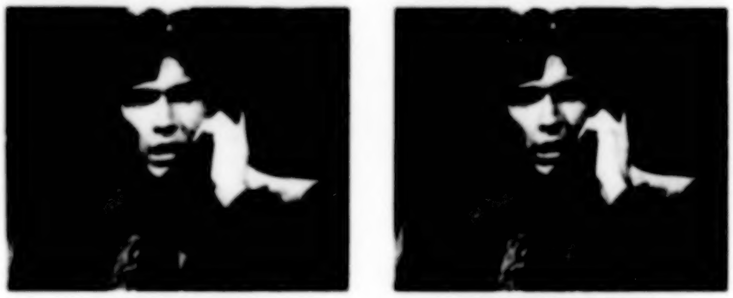
SCENE	BITS/PIXEL
BEACH	2.228
MAKE UP	1.738
LAWYER	1.976
STAR TREK	1.682
BEES & FLOWERS	1.996
GAME/GIRL	1.689
WOMAN/COUCH	1.813
HOSPITAL	1.979
WOMAN/MAN	1.802
WOMAN/GRAY SUIT	1.872
TWO MEN	1.949
MAN/PHONE	1.815
FULLER BRUSH	2.037
PLANE/BI-PLANE	1.711
WOMAN/HAND	1.671
NEWS WOMAN	1.823
COLOR BARS	1.347
TWO WOMEN	1.659
AVERAGE	1.822

ORIGINAL PAGE IS
OF POOR QUALITY

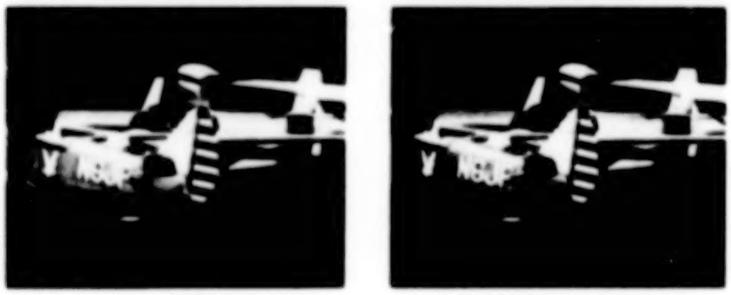
ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



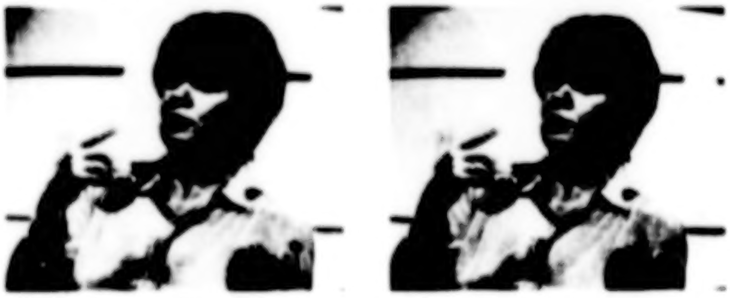
(a) Original image (8 bpp). (b) Processed image (1.849 bpp).



(c) Original image (8 bpp). (d) Processed image (1.815 bpp).



(e) Original image (8 bpp). (f) Processed image (1.711 bpp).



(g) Original image (8 bpp). (h) Processed image (1.671 bpp).

Figure 2. - Results of computer simulation processing of NTSC video images (bpp – bits per pixel).

N91

24068

UNCLAS

N91-24068

Recent Advances in Coding Theory for Near Error-Free Communications

K-M. Cheung, L.J. Deutsch, S.J. Dolinar, R.J. McEliece, F. Pollara, M. Shahshahani, L. Swanson

Communications Systems Research Section – Jet Propulsion Laboratory [*]

1 Introduction

In 1948 Claude Shannon revolutionized communication technology forever with the publication of his classic paper "A mathematical theory of communication." In this paper he wrote

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point."

He then showed that the most efficient solution to this fundamental problem necessarily involved *coding* the selected message, before transmission over the channel. Shannon did not say exactly how this coding should be done; he only proved mathematically that efficient coding schemes must exist. Since 1948, a whole generation of later researchers has validated Shannon's work by devising explicit and practical coding schemes, which are now part of practically every modern digital communication system. Near-earth satellite communication systems, high performance military systems, computer communication networks, high speed modems, and compact-disk recording and playback systems, all rely heavily on sophisticated coding schemes to enhance their performance.

The communication channel which has benefitted most from Shannon's insights, perhaps, is the *deep-space channel*, i.e., the channel by which spacecraft like *Pioneer*, *Mariner*, *Voyager*, etc., transmit their astonishing views of the Solar System back to planet Earth. In this paper we will discuss the benefits that coding has already brought to deep-space communication, and the benefits that may yet become a reality.

Coding has traditionally been divided into two branches: *channel coding* and *source coding*. The goal of channel coding is to protect the transmitted message from the errors and distortions which may be caused by the channel. The goal of source coding (sometimes also called *data compression*), on the other hand, is to ensure that the transmitted message is as dense with information as possible. While channel coding has historically been the more important contributor to deep-space communication technology, future gains may come largely from source coding.

2 Channel Coding

Space communication engineers realized, almost from the beginning of the space age, that Shannon's ideas could be used to improve the design of spacecraft telemetry systems. Scarcely 10 years after the launch of the primitive *Explorer* spacecraft in 1958, NASA had begun the routine use of channel coding to enhance deep-space communications. In 1968, the *Pioneer 9* solar orbiter was launched with an encoder for a rate 1/2, constraint length 25 convolutional code to be decoded by the Fano sequential decoding algorithm. Similar encoders and decoders were used on later *Pioneers* which explored Jupiter, Saturn, and Venus. In 1969, *Mariners VI and VII*, twin Mars flyby missions, were launched with a (32, 6) biorthogonal code which enhanced their performance by 2.2 dB at a decoded bit error probability of 5×10^{-3} (all performance gains in this section are compared to uncoded data.) This biorthogonal code was also used on the 1971 *Mariner 71* Mars orbiter and the 1975 *Viking* Mars orbiter and lander. Then in 1977 with the launch of the *Voyager I and II* spacecraft, a new generation of deep-space channel coding began. *Voyager* used a rate $r = 1/2$, constraint length $K = 7$ convolutional code, which was decoded on the ground using Viterbi's revolutionary decoding algorithm. This system achieved a performance gain of 3.5 dB at a decoded bit error probability of 5×10^{-3}. The coding system on Voyager was further enhanced by an 8-bit (255, 223) Reed-Solomon code, which was used in concatenation with the convolutional code to produce the low bit error rates needed by the data compression scheme described in Sec. 3.1. This coding system, which has become a standard for NASA and its Deep Space Network (DSN), is also in use on the *Galileo*, *Magellan*, and *Ulysses* missions. However, there is more to the *Galileo* coding story, as we will see in the following section.

---

[*] This article represents the results of one phase of research carried out at the Jet Propulsion Laboratory, California Institute of Technology, sponsored by the National Aeronautics and Space Administration.

Other significant new results on efficient coding for severely bandlimited channels, as in the mobile satellite environment, have been developed at JPL [5], but cannot be described in detail in this article.

2.1 Large Constraint Length Convolutional Codes: The *Galileo* Code

As mentioned in the previous section, the $K = 7$, $r = 1/2$ convolutional code, enhanced at low decoded bit error probabilities by concatenation with a (255, 223) Reed-Solomon code, has been a NASA standard since the launch of *Voyager* in 1977. However, according to Shannon's theorems, the performance of this system could in principle be improved by a further 4dB at a decoded bit error probability of 5×10^{-3}. Motivated by this tantalizing fact, in 1982, DSN Advanced Systems undertook a long-term research effort to study advanced coding techniques which would yield significantly more coding gain than is available using the NASA standard codes. Since Shannon had shown that as one approached his theoretical limits, the implementational complexity would increase explosively, the specific target of this research was a 2dB improvement, about half of the maximum possible theoretical gain. The hope was that this research effort would yield improved coding systems for missions in the "far future."

The quest for the 2 dB coding gain took off in several directions from current codes. The research focused on the same basic concatenation of a Reed-Solomon outer code with a convolutional inner code, but the code parameters were allowed to vary to levels not feasible when the present NASA standards were developed. The research effort studied the effects of increasing the constraint length and decreasing the code rate of the convolutional code, and increasing the symbol size and optimizing the code rate of the Reed-Solomon code. Due to a higher predicted payoff in performance versus complexity, a significant advance in convolutional code parameters was attempted, whereas the Reed-Solomon code parameters were only varied slightly from those of the present Reed-Solomon code used on *Voyager* and *Galileo*.

In 1986 the quest was declared a success, when, after extensive computer searches, some codes were found which surpassed the 2 dB goal, the best code improving performance by 2.11 dB (See Fig. 1 for a comparison of several coding systems.) To achieve this gain, the convolutional code constraint length had been increased to $K = 15$ and the code rate decreased to $r = 1/6$, and a 10-bit (1023, 959) Reed-Solomon code was used as the outer code [11].

Then on January 28, 1986 the *Challenger* catastrophe caused a three-and-a-half year postponement of *Galileo*'s launch. During this delay, a search was conducted for an advanced convolutional code which could be used by *Galileo* as an experimental mission enhancement option. Luckily, the 2dB code search was by then complete, and its results were made available to the *Galileo* team. The actual $K = 15$, $R = 1/6$ code could not be used by *Galileo*, since the bandwidth of *Galileo*'s radio modulator limits code rate to be at least 1/4. However, a fairly simple modification of the search that lead to the "2dB" code led to the discovery of a $K = 15$, $r = 1/4$, which was adopted for use by the mission. Because of *Galileo*'s bandwidth constraints, this code's gain falls short of the full 2dB.

Still, the *Galileo* code, as it is now called, will realize between 1 and 2dB gain beyond the standard NASA code, which may significantly improve the science return for this important mission. When *Galileo* was finally launched in October of 1989, it carried with it an experimental encoder for the new code, which incidentally also required a mechanism for doubling the subcarrier frequency and symbol clock rate whenever the encoder is invoked. Thus did the results of a research effort aimed at the "far future" affect a mission in the immediate present! [6]

However, at launch time, no decoder for the experimental code existed. The design of the decoders for such codes, suitable for VLSI implementation, is a current research project at JPL. The evolution and the results of this effort are the subject of the discussion in the following section.

2.2 Decoder Design: The Big Viterbi Decoder

The higher coding gains and the ensuing higher information returns promised by new, powerful coding schemes are becoming a reality through the design and implementation of the complex decoders that are required to take full advantage of such codes.

In particular, the complexity of a Viterbi decoder depends on three main parameters: the constraint length K, the code rate r, and the information data rate. The major complexity driver is the constraint length, since the amount of hardware is increasing exponentially with K. A decoder for $K = 15$ has 2^{14} states and is approximately 256 times more complex than the decoder for the $K = 7$ Voyager code.

Such decoders cannot be implemented on a single VLSI chip, because of memory storage and throughput rate requirements. Therefore, one is forced to develop concurrent algorithms which can be used on a multiprocessor system. The Viterbi algorithm is inherently a parallel algorithm. However, a fully parallel implementation of a large Viterbi decoder requires an impractical amount of hardware. This leads to the fundamental question of how to efficiently exploit this parallelism and how to contain it into practical limits by introducing some sequential re-use of the available hardware, which inevitably reduces the overall decoding speed.

Extensive research performed at Caltech and at JPL on multi-processor computers consisting of a network of simple node-processors interconnected as a n-dimensional cube (Hypercube), suggested the possibility to implement the Viterbi algorithm on a hypercube computer. The Hypercube or n-dimensional cube is a natural topology for efficient implementation of the Fast Fourier Transform (FFT). The similarity between the Viterbi algorithm and the FFT, which can be described in terms of the same graph topology and different algebraic kernels, was exploited to show how a network of 2^n processors, interconnected as a n-dimensional cube with no shared memory, can be efficiently used to implement a Viterbi decoder with various degrees of parallelism. The interprocessor communication overhead can be reduced by using the traceback method [4], which completely avoids any survivor exchange or any global memory operation. This algorithm has been implemented and successfully tested on a 64-node general purpose Hypercube computer for $(K, 1/N)$ codes, with $K = 3, ..., 15$, with measured efficiencies as high as 65 % for $K = 15$.

After this first experience with parallel decoding algorithms, it was decided to develop a Viterbi decoder, capable of decoding convolutional codes with constraint length up to 15 for NASA's Deep Space Network. Prototypes of the Big Viterbi Decoder (BVD) are being built both in semi-custom VLSI and gate array technology, and will use 512 or 256 identical VLSI chips, respectively, distributed on 16 identical boards. The final implementation will probably use a single board populated by 64 VLSI chips. The projected data rate will be approximately 1 Mbit/sec which is well in excess of present mission requirements. Galileo, for example, sends telemetry data at a maximum rate of 134.4 Kbit/sec.

The decoder could be implemented using serial or parallel architectures, or with a hybrid serial-parallel approach. In a serial architecture, a single physical butterfly processor performs all 8192 butterflies, sequentially. In a hybrid approach, n physical butterflies are used, each sequencing through $8192/n$ butterflies.

A fully parallel architecture was chosen for the BVD, since it is more convenient, in terms of VLSI implementation, to sacrifice parallelism by using bit-serial arithmetic. Therefore, additions and comparisons of metrics are done bit-serially, and the results are sent serially to other butterflies.

The most challenging problem in the design of the decoder is the wiring of the chips. This wiring is based on a novel partitioning [3], [10] of the decoder's state transition diagram and it defines the new decoder's architecture.

When the constraint length K is large, it is desirable to take a modular, hierarchical approach to organizing the huge number of required elements. Many add-compare-select circuits can be implemented on a single VLSI chip, and many chips can be mounted on a single printed circuit board. The full decoder is implemented by wiring together the required number of chips and boards.

The connection diagram of the 2^{K-2} butterflies is a deBruijn graph [7]; the butterflies are nodes in the graph and the edges of the graph represent wires between butterflies. It is possible to split the set of butterflies into modules called boards and the boards into modules called chips, in such a way that a large proportion of the required connections between butterflies are implemented internally within the modules. The chips are all identical (see Fig. 2) and the boards are all identical. Furthermore, their internal structure does not depend on the size of decoder, and an appropriate number of these *universal* modules can be wired together to make any size decoder.

Consider as an example a $K = 15$ Viterbi decoder consisting of 16 boards and 512 chips. Each chip in this design contains 16 butterflies, and each board has 32 chips. However, the theory developed in [3] is completely general and produces a modular, hierarchical partitioning of any size deBruijn graph into any number of first-level and second-level subgraphs (boards and chips), according to an FFT-like connection pattern.

The total number of wires cannot be increased or reduced by any wiring scheme. However, it is advantageous to capture as many of these required connections as possible within identical, small, modular units (chips and boards). The board and chip modules defined by this FFT-like construction have the property that full Viterbi decoders of all sizes can be constructed by appropriately connecting identical copies of the universal module, without revising the internal wiring within any module.

The success of the Galileo coding experiment depends now on the implementation of the $K = 15$ Big Viterbi Decoder, which should be completed by late 1990.

3 Source Coding

Although the theory of source coding is almost as old as that of channel coding, applications to deep-space imaging have been much slower to develop. There are several reasons for this. First, it is quite difficult to produce credible mathematical models for image statistics — this is in sharp contrast to deep-space channel modeling, where the Gaussian model has served admirably for 30 years or more. Second, and perhaps more important, in source coding the burden of complexity is on the *encoders*, which must be on the spacecraft and so highly constrained in power, weight, etc., whereas in channel coding, the burden is on the *decoder*, which is on the ground and so relatively unconstrained. A further complication is the difficulty in formulating a meaningful distortion measure. Nevertheless, beginning with *Voyager's* 1986 encounter with Uranus, source coding has begun to play an important part in deep-space telecommunications systems.

3.1 Voyager's and Galileo's Data Compression Scheme

The *Voyager* mission, at Uranus in 1986 and Neptune in 1989, used a source coding technique pioneered by R. F. Rice, and achieved a compression ratio of 2.5:1. This algorithm is essentially a universal source code on the differences between successive pixels. An enhanced version of Rice's algorithm, known as the BARC (block adaptive rate controlled) scheme will be used on *Galileo's* images of Jupiter [9]. A general version of Rice's algorithm is now being developed for NASA in VLSI by the University of Idaho.

However, both of Rice's schemes (Voyager and BARC) were constrained by the need to use 1970's space-qualified hardware. With the availability of enormously more powerful VLSI hardware in the

1990s and beyond, it is now possible to envision a future in which source coding can deliver gains to deep-space telemetry fully comparable to those already realized by channel coding.

3.2 Current Research in Data Compression for Images

The theoretical limits of combined source-channel coding for a Gauss-Markov source and a Gaussian channel are illustrated in Fig. 3, where the performance is measured in terms of mean square reproduction error as a function of the signal-to-noise ratio of source symbols, for two values of bandwidth expansion factor $\eta = \infty$ and $\eta = 4$. The correlation coefficient γ represents the dependency between successive pixels. No communication system can be designed below the curves for $\eta = \infty$. For a practical bandwidth expansion factor ($\eta = 4$), it is shown how the exploitation of the spatial redundancy of a typical planetary image (with correlation coefficient $\gamma = 0.99$) offers a potential improvement of approximately 5 dB. These large potential gains motivate the current research in source coding.

In the area of lossless compression schemes, we have developed a method based on traditional DPCM (Differential Pulse Code Modulation) encoded with a GVH (Gallager-vanVoorhis-Huffman) code, which is a near optimal adaptive Huffman code using simple table look-up operations, and requiring very modest computation [1].

However, lossy compression schemes are the most promising for achieving large gains. The same GVH code has been applied to the coefficients resulting from a two-dimensional discrete cosine transform (DCT) on 8 × 8 or 16 × 16 blocks of pixels. This strategy yields a scheme with negligible distortion and high compression ratios (10:1 to 20:1 on typical images). This technique is based on the fact that the DCT coefficients are approximately geometrically distributed and the GVH code is optimal in this case.

For cases where higher distortion is allowable, a new locally adaptive vector quantization scheme has been developed with compression ratios in the range 30:1 to 60:1. This is a new class of source coding schemes based on simple dynamic codebook updating strategies, and on self-organizing data structures [2]. This scheme does not assume any a priori knowledge of the source statistics, nor does it require any preprocessing or codebook training. The codebook is generated on the fly, and is constantly updated to capture local features of the data.

4 Future Research Directions

4.1 Neural Networks for Soft Decoding

Block codes are traditionally decoded by first "hard quantizing" symbols, i.e. by replacing soft channel symbols by zeros and ones, followed by algebraic decoding. This has the disadvantage of losing about 2dB compared to soft, maximum-likelihood decoding. The problem of soft decoding of a (n, k) block code consists in finding the distances of a n-dimensional vector of soft channel symbols to the 2^k codewords, and in selecting the closest codeword. Unfortunately, known soft maximum-likelihood decoding methods are extremely complex, and so block codes have rarely been used as inner codes for deep-space applications, where the 2dB loss is too large.

Recent results obtained by JPL researchers show that neural networks can be efficiently used to implement such decoders, yielding highly parallel structures and extremely high decoding speeds. The decoding problem can be cast in a form suitable for multi-layer feed-forward neural networks (perceptrons). This can be accomplished by observing that the decision regions for the 2^k codewords are convex regions delimited by hyperplanes in n-dimensional space, and multi-layer networks can be configured to describe exactly such regions. Two layer networks are sufficient to solve this problem, while the conventional search for the minimum distance codeword implies $\log_2 2^k = k$ sequential stages or layers. Thus very fast decoding should be possible.

The number of neurons necessary for such decoders was analyzed together with the performance penalties incurred by reducing such number. The fault tolerance properties inherent in neural networks could be valuable in on-board or remote applications. Learning rules (the back propagation rule, etc.) allowing efficient use of the available neurons by finding the optimal weights to be assigned to the branches of the network are being developed.

The greatest potential of neural nets decoders is the high-speed processing that could be provided through massively parallel VLSI implementations.

4.2 Neural Networks for Source Coding

Recent research conducted at JPL has shown how a vector quantizer can be used for image compression by mapping a sequence of continuous or discrete vectors, representing a rectangular block of pixels, into a lower rate sequence, suitable for communication over bandwidth constrained digital channels.

Traditional vector quantizer encoders are computationally very intensive. Highly parallel feed-forward neural networks can be used to implement a class of memoryless vector quantizer encoders, with encoding time proportional to $\log_2 N$, in contrast to the KN of traditional encoders, where N is the codebook size and K is the dimension of each vector.

Feed-forward neural network implementations of a class of finite-state vector quantizer encoders are also being developed to take advantage of the correlation between successive source vectors.

It was estimated that these vector quantization schemes can provide a 10:1 compression ratio with little distortion, and a 100:1 compression ratio with moderate distortion for several sources. The squared error distortion measure is used for comparing different schemes, together with other subjective distortion measures.

Theoretical performance results were verified by simulation, and a preliminary design and architecture for VLSI implementation of the proposed compression systems is under development.

4.3 Finite-state Codes

Methods for obtaining large coding gains using a new class of hybrid trellis codes have been developed at JPL. These codes are constructed by combining block with convolutional codes in a novel manner, using the recent idea of "set partitioning" developed by Ungerboeck and others. We believe that these codes can significantly improve the performance of coded communication systems without requiring a significant increase in decoder cost.

A convolutional code is characterized by several integer parameters: k, the number of information bits that enter the encoder at each clock cycle; n, the number of transmitted bits that leave the encoder at each clock cycle; and m, the memory, which represents the number of previous k-bit information blocks that the current n-bit output block depends on. In principle, these parameters can assume any integer values, but for historical (and other) reasons almost all current implementations of convolutionally encoded systems use very small values for k and n, and rely on the encoder memory m to obtain the required coding gains. (For example, the convolutional code used by Voyager has $k = 1$, $n = 2$, and $m = 6$.) However, there are convincing theoretical arguments that show the existence of powerful and practical convolutional codes for which the parameters n and k are relatively large, while the memory m is relatively small. Unfortunately, it has always been very difficult to locate such codes because there is no really satisfactory algebraic description of them and brute-force computer searches are prohibitively difficult.

However, we have recently discovered a large class of convolutional codes with large n and k, and small m, which can be described algebraically, and which our preliminary studies have shown to contain some very powerful codes [8]. Our basic idea is as follows. We choose the parameters n and k to correspond to the block length and dimension, respectively, of a known block code. We then partition this block code into a number of smaller subcodes, and then assign codewords from these subcodes to the branches of the trellis diagram corresponding to a convolutional code with small memory, as shown in Fig. 4. The codes constructed this way can have higher coding gain than the original block code, increased ability to correct short bursts of errors, and less decoder complexity than traditional convolutional codes with the same level of performance. "Higher coding gain" will translate into lower decoded bit error probability; the short burst error correcting property will make these codes attractive as inner codes in concatenated coding systems, and the decreased complexity will make VLSI implementation feasible.

Additional advantages include the existence of higher rate codes (smaller bandwidth expansion) with good performance, and the possibility to avoid lengthy computer searches to find good codes.

4.4 Fractals for Data Compression

The use of fractals has been proposed to capture very complex images in a few parameters. This has the potential for huge compression ratios, at the expense of very complex encoding algorithms and moderate distortion.

The idea of the fractal method is to identify a data set with the stationary distribution of a Markov chain determined by a finite set of affine transformations. While a Markov chain can be specified by very few parameters, its stationary distribution may indeed be a very complex object. The goal is to take advantage of this fact and obtain concise representations of very complex objects or data sets by simply encoding the parameters of the Markov chain. Successful development of the technology for the implementation of this idea would have very significant implications for source coding.

While many images may be generated by the above mentioned method, the procedure requires human supervision and is very time consuming. Therefore the key issue is to systematize the procedure so that it can be implemented in an automated fashion. We have had some success with such a method for one dimensional data. In two dimensions, for instance for images, we face a new set of problems due to the complexity of the data. It has been suggested that by using the method of "skeletonization" one can determine the parameters of the Markov chain whose stationary distribution is the given image. We are presently experimenting with this idea.

References

[1] K. Cheung, P. Smyth and H. Wang, "A High-speed Distortionless Predictive Image-compression Scheme", TDA Prog. Rep. 42-102, JPL, Pasadena, Aug. 15, 1990.

[2] K. Cheung and V.K.Wei, "A Locally Adaptive Source Coding Scheme", submitted to IEEE Trans. on Communication Theory.

[3] O. Collins, F. Pollara, S. Dolinar and J. Statman, "Wiring Viterbi Decoders (Splitting DeBruijn Graphs", TDA Prog. Rep., JPL, Pasadena, Feb. 15, 1989, pp. 93-103.

[4] O. Collins and F. Pollara, "Memory Management in Traceback Viterbi Decoders", TDA Progress Report 42-99, Jet Propulsion Lab., Pasadena, CA, Nov. 1989, pp. 98-104.

[5] D. Divsalar and M.K. Simon, "The Design of Trellis Coded MPSK for Fading Channels", IEEE Trans. on Comm., Vol. 36, No. 9, pp. 1013-1021, Sept. 88.

[6] S. Dolinar, "A New Code for Galileo," TDA Prog. Rep. 42-93., May 15, 1988, pp. 83-96.

[7] S.W. Golomb, "Shift Register Sequences", California: Aegean Press, 1982.

[8] F. Pollara, R. McEliece, and K. Abdel-Ghaffar, "Finite-state Codes," IEEE Trans. Inform. Theory, vol. 34, No. 5, Sept. 1988.

[9] R.F. Rice and J. Lee, "Some Practical Universal Noiseless Coding Techniques, Part II", JPL Publ. 83-17, JPL, Pasadena, March 1, 1983.

[10] J. Statman, G. Zimmerman, F. Pollara and O. Collins, "A Long Constraint Length VLSI Viterbi Decoder for the DSN", TDA Prog. Rep. 42-95, JPL, Pasadena, Nov. 15, 1988, pp. 134-142.

[11] J.H. Yuen and Q.D. Vo, "In Search of a 2 dB Coding Gain", TDA Prog. Rep. 42-83, JPL, Nov. 85, pp. 26-33.

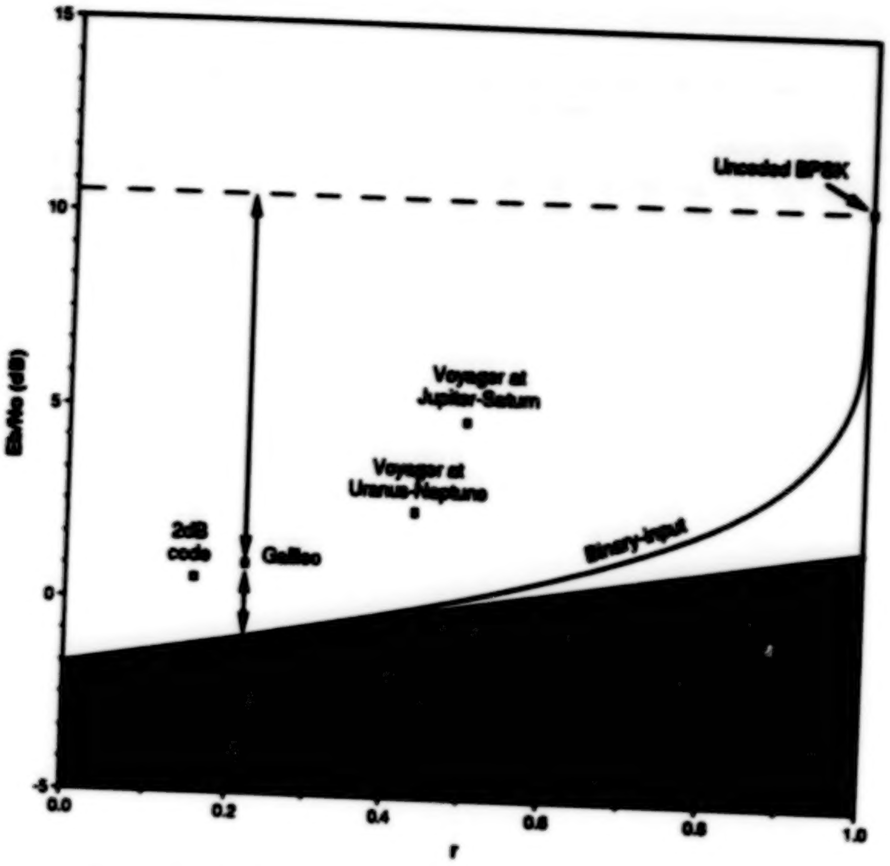


Figure 1 Performance comparison of several coding systems.

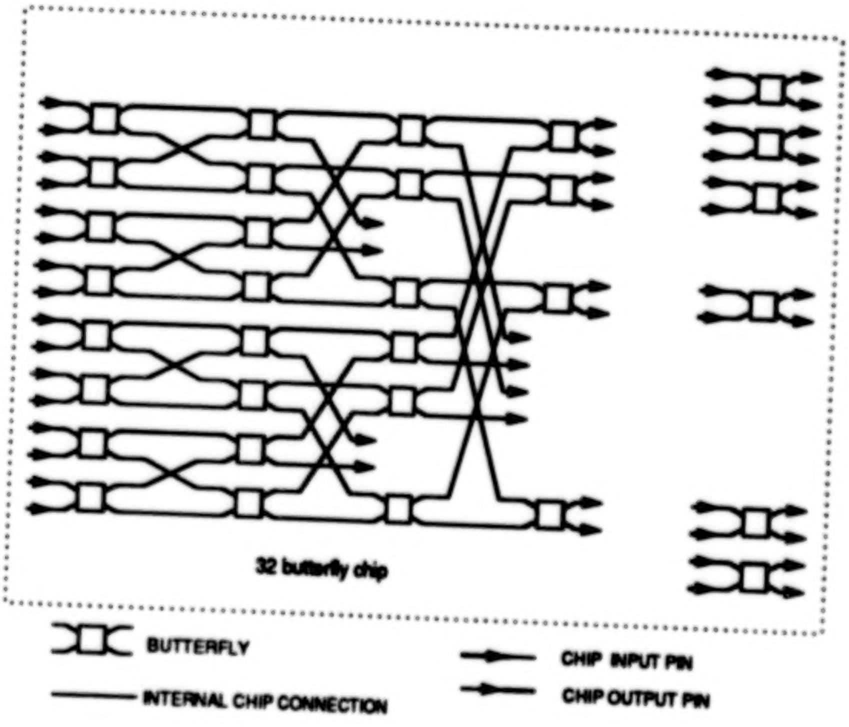

32 butterfly chip
BUTTERFLY
INTERNAL CHIP CONNECTION
CHIP INPUT PIN
CHIP OUTPUT PIN

Figure 2 Example of universal BVD chip.

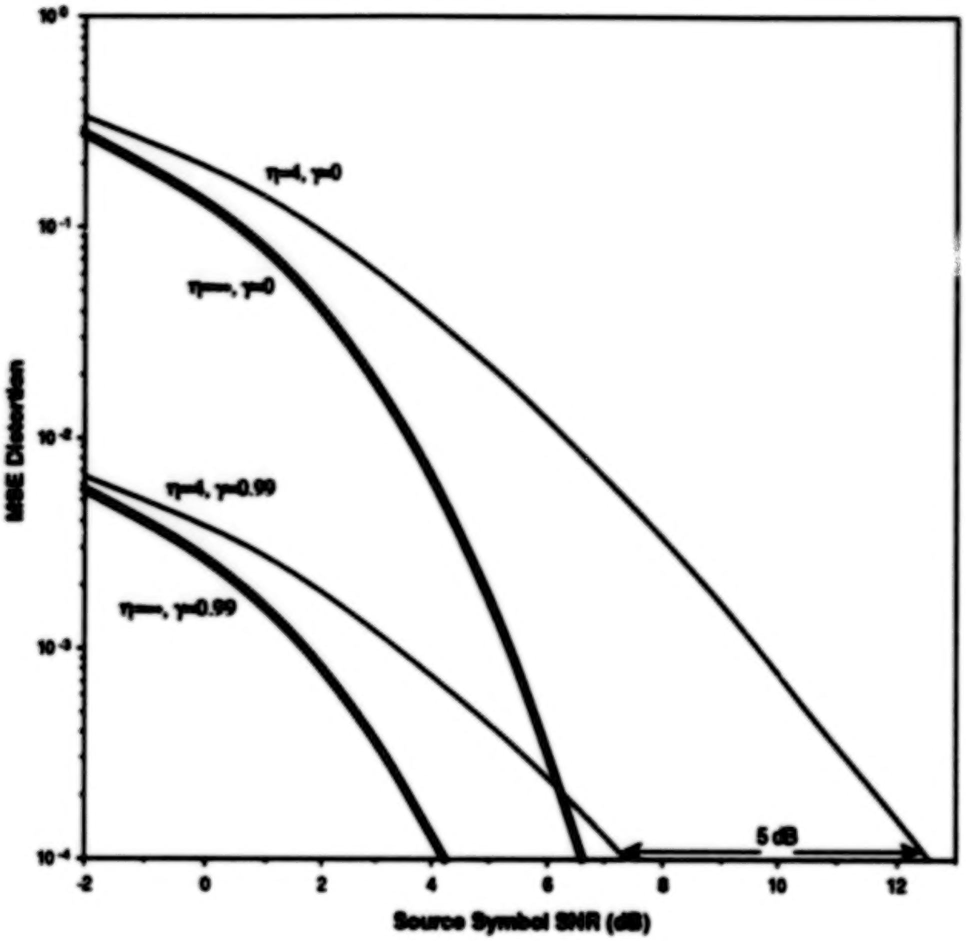
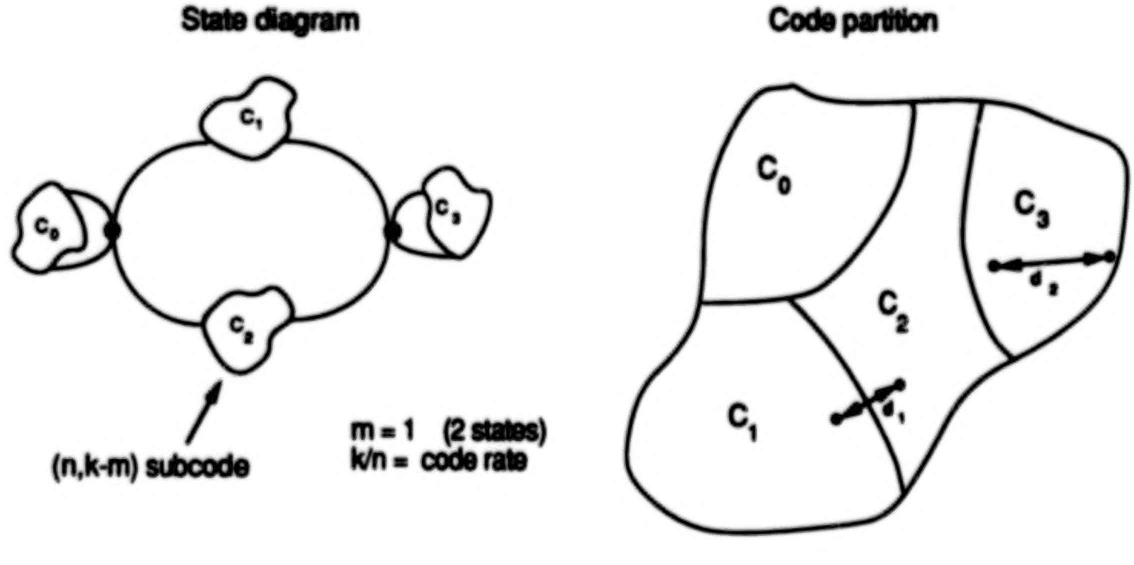


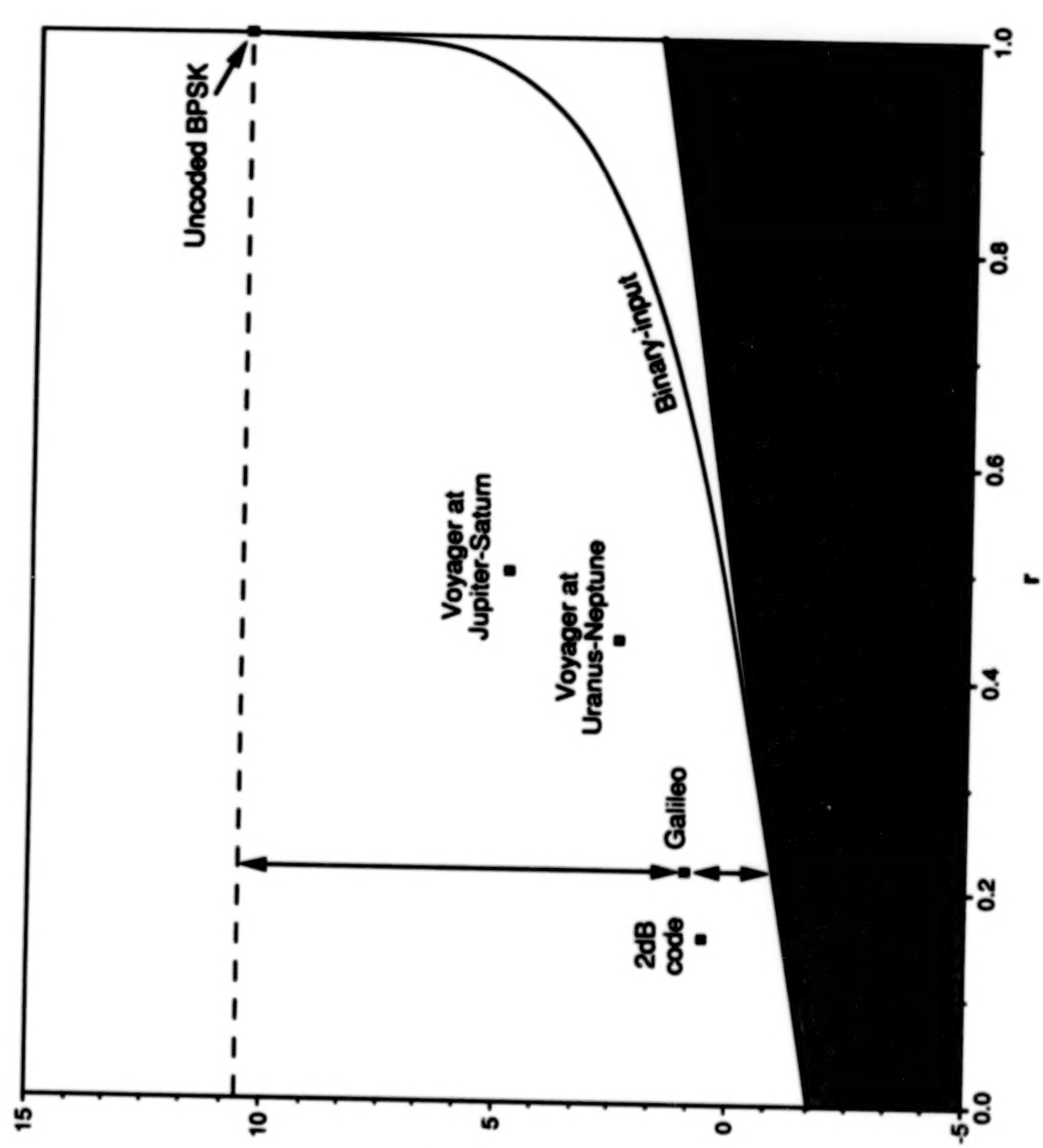
Figure 3 Performance gains of source coding

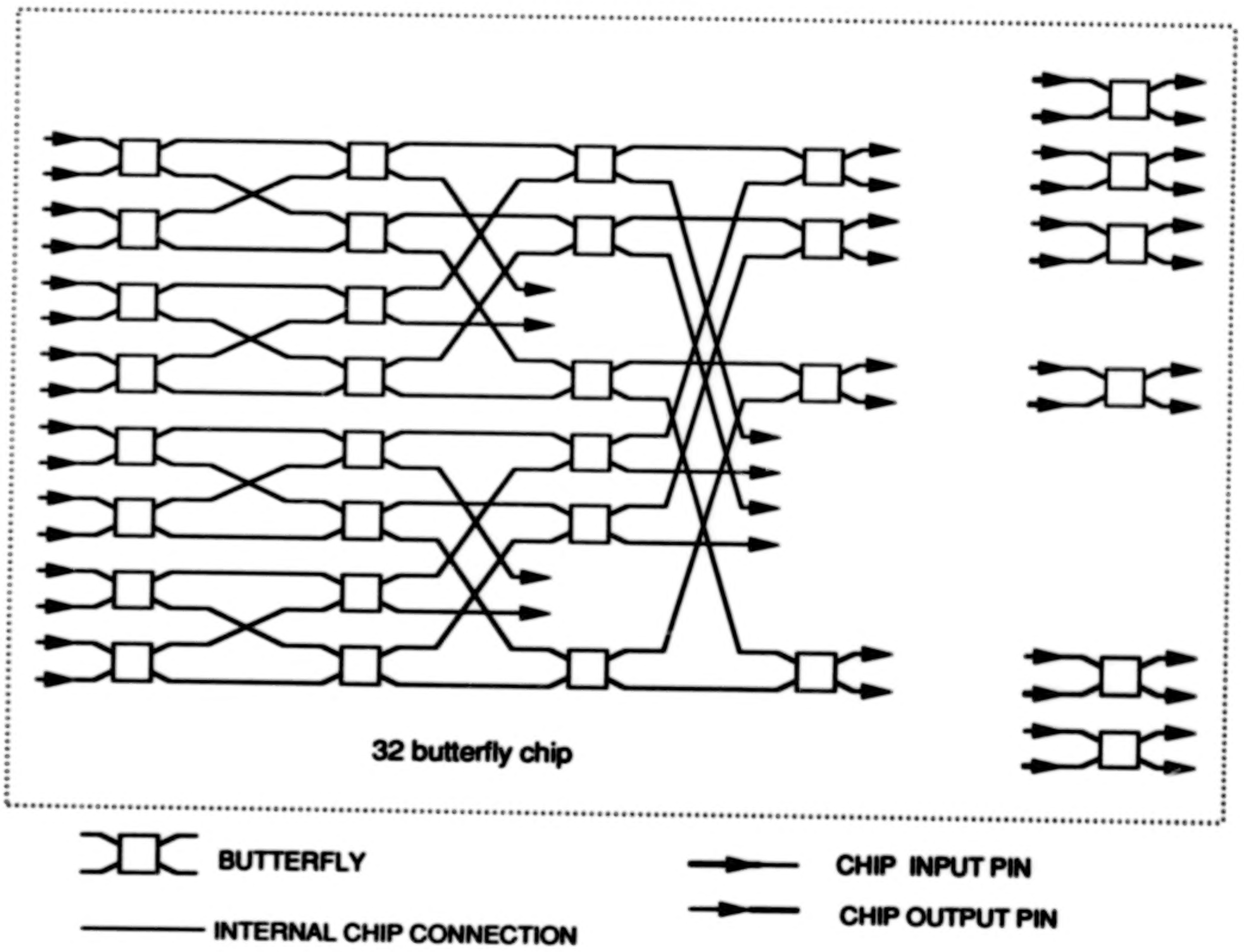

State diagram

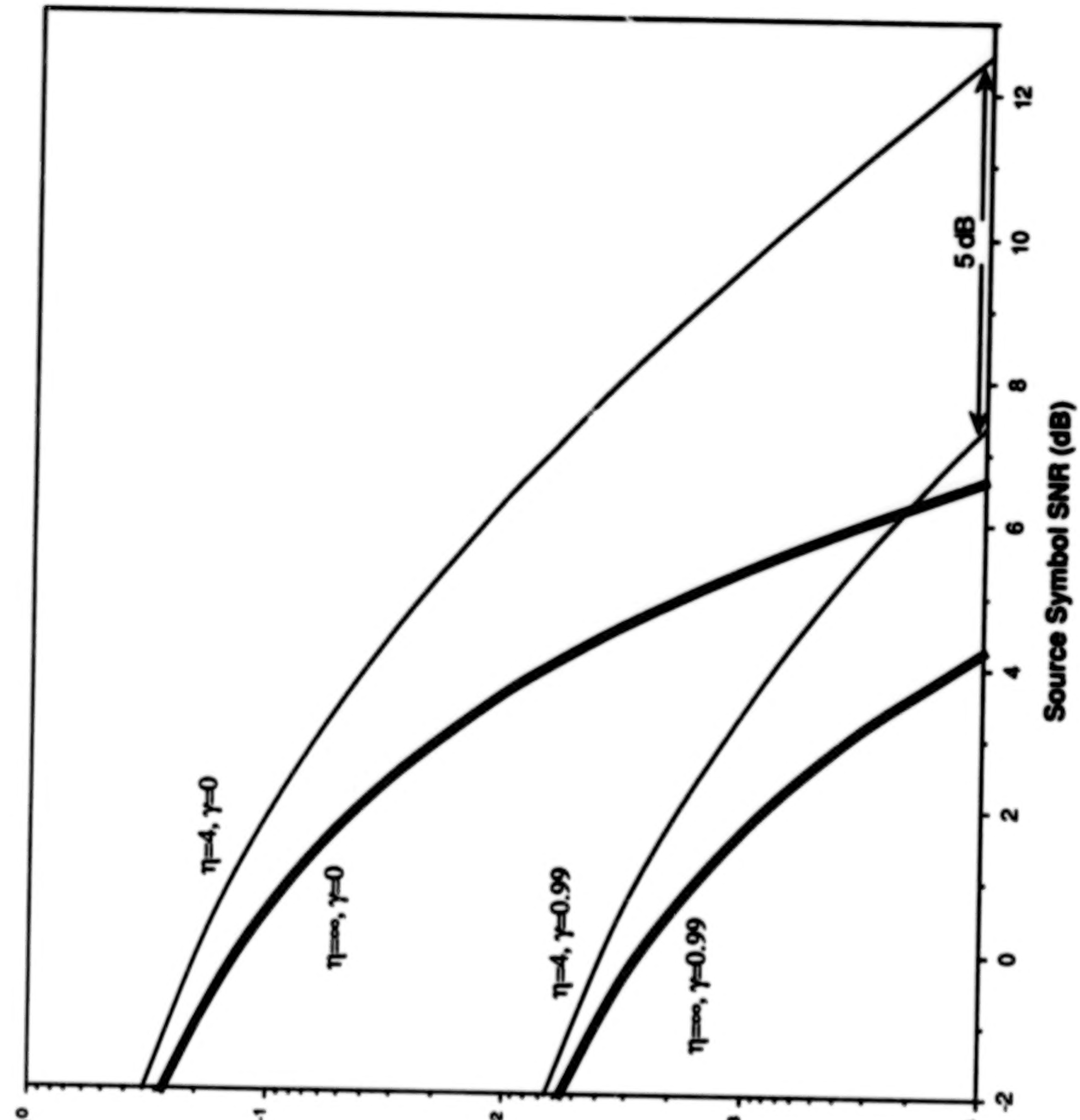
Code partition

C_1
C_0
C_3
C_2
(n,k-m) subcode
m = 1 (2 states)
k/n = code rate
C_0
C_3
d_2
C_2
C_1
d_1

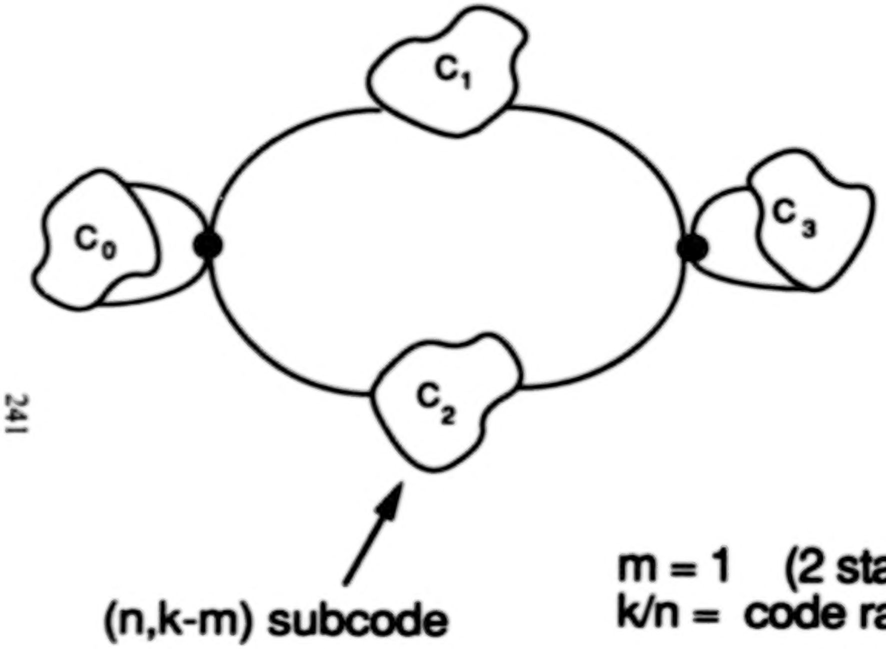
Figure 4 Example of finite-state code


Eb/No (dB)
15
10
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0
r
Uncoded BPSK
Binary-input
Voyager at Jupiter-Saturn
Voyager at Uranus-Neptune
Galileo
2dB code


32 butterfly chip
BUTTERFLY
INTERNAL CHIP CONNECTION
CHIP INPUT PIN
CHIP OUTPUT PIN

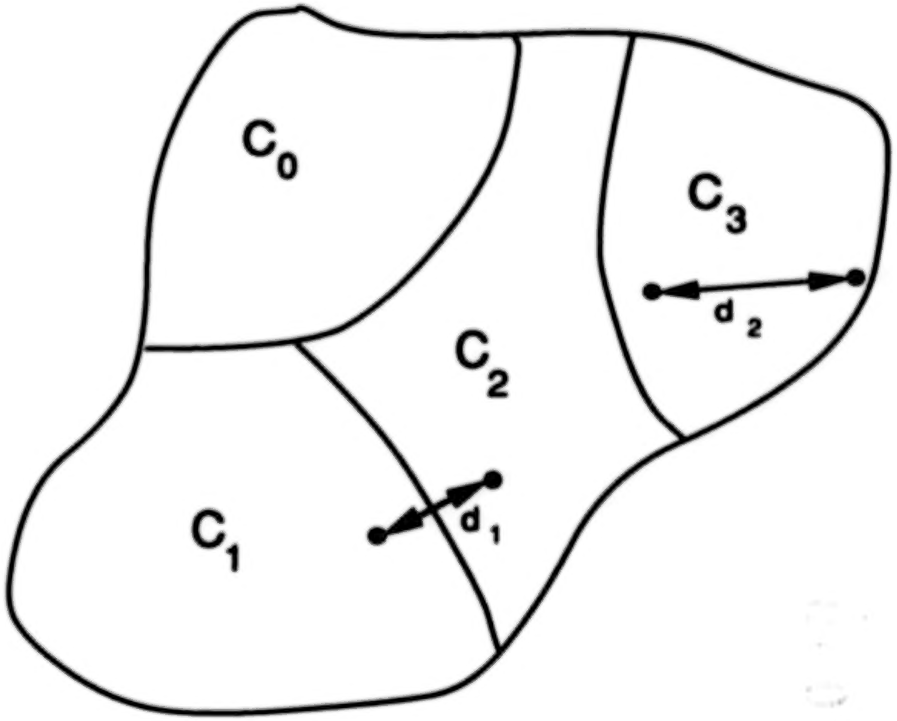

MSE Distortion
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
Source Symbol SNR (dB)
-2
0
2
4
6
8
10
12
$\eta=4,\ \gamma=0$
$\eta=\infty,\ \gamma=0$
$\eta=4,\ \gamma=0.99$
$\eta=\infty,\ \gamma=0.99$
5 dB

State diagram


C₁
C₀
C₃
C₂
(n,k-m) subcode

m = 1 (2 states)
k/n = code rate

Code partition


C₀
C₃
d₂
C₂
C₁
d₁

N91

24069

UNCLAS

N91-24069

MICROWAVE INTEGRATED CIRCUITS FOR SPACE APPLICATIONS

Regis F. Leonard and Robert R. Romanofsky

NASA Lewis Research Center
21000 Brookpark Road
Cleveland, Ohio 44135

Monolithic microwave integrated circuits (MMICs), which incorporate all the elements of a microwave circuit on a single semiconductor substrate, offer the potential for drastic reductions in circuit weight and volume and increased reliability, all of which make many new concepts in electronic circuitry for space applications feasible, including phased array antennas.

Over the last ten years, NASA has undertaken an extensive program aimed at development of MMICs for space applications. The first such circuits targeted for development were an extension of work begun earlier in hybrid (discrete component) technology in support of the Advanced Communication Technology Satellite (ACTS). As a result it focused on power amplifiers, receivers and switches at ACTS frequencies. More recent work, however, has focused on frequencies appropriate for other NASA programs and emphasizes advanced materials in an effort to enhance efficiency, power handling capability, and frequency of operation or noise figure to meet the requirements of space systems.

MONOLITHIC MESFET CIRCUITS

Background

This work has been carried out under contract with Texas Instruments, Rockwell, and Minneapolis Honeywell and Hughes. It is aimed at ACTS-like applications, namely, communications satellites which feature multiple, electronically steerable beams. Such a system would ideally be implemented using phased array antennas with lightweight, low volume distributed transmitters and receivers and lightweight monolithic phase shifters rather than the extensive network of ferrite phase shifters and switches used by ACTS. The initial stages of development therefore were aimed at the development of the monolithic circuitry required for such a system. Carried out between approximately 1983 and 1987 and featuring transmitter frequencies of 20 GHz and receiver frequencies of 30 GHz, as required for the satellite portion of the system, the program utilized as a basic device the GaAs metal/semiconductor field effect transistor (MESFET). The program produced a number of developments, each of which consisted of one or more monolithic chips. Some of these are described here.

Ku-Band Amplifier

The space station proximity communications system is intended to provide communications within a radius of several 10's of kilometers of the space station. Users would include free flying experimental platforms, the orbital maneuvering vehicle (OMV), and astronauts in EVA. The original planning for this system utilized frequencies in Ku-band. It is not clear at this time whether that assignment will be maintained, inasmuch as there exists possible interference with commercial, fixed satellite services. Nevertheless, several chips have been developed at 13-15 GHz to accommodate this application. The most challenging of these was a variable power amplifier intended primarily for use by astronauts during EVA. The system design requires approximately 1 watt of output power. Power variability is necessary because of the wide variation in range experienced by the astronaut. Of course, high efficiency is also a prime consideration. The development of the chip was undertaken by Texas Instruments in January 1987. Their design is a four stage monolithic circuit using dual gate GaAs MESFET technology. The chip is shown in Figure 1.

ORIGINAL PAGE IS
OF POOR QUALITY

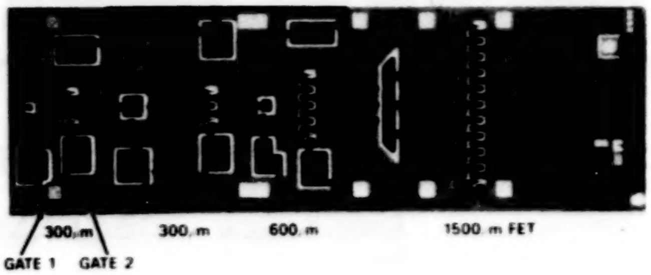


FIGURE 1. 15 GHz MONOLITHIC VARIABLE POWER AMPLIFIER

The performance of the amplifier, compared to the design goals, is shown in Table 1.

TABLE 1.
15 GHz HIGH EFFICIENCY VARIABLE POWER AMPLIFIER

	DESIGN GOAL	PERFORMANCE
CENTER FREQ./BANDWIDTH (GHz)	2.0/14.0	2.0/17.0
GAIN (dB)	15.0	29.6
MAX POWER OUTPUT (W)	1.0	.92
EFFICIENCY AT MAX POWER (%)	35	30
FINAL STAGE GATE WIDTH (MM)	---	1.5

20 GHz Variable Power Amplifier

Since, ideally, a phased-array antenna should be able to vary the phase and the amplitude of each antenna element independently, one desirable module for such a system would be a variable power amplifier. Because of the stringent limitations on power consumption imposed by a space system, it is desirable to maintain insofar as possible the efficiency of a power amplifier while adjusting output power. For this reason, the contractor (Texas Instruments) chose a dual gate FET as the basic device for design of this chip, and adjusted the device bias to obtain variable output. The topology of the amplifier closely resembles the Ku-band amplifier illustrated earlier. This was one of the earliest MMICs developed under the NASA program. A summary of its performance compared to design goals is shown in Table 2.

TABLE 2
TEXAS INSTRUMENTS 20 GHz VARIABLE POWER AMPLIFIER

	DESIGN GOAL	PERFORMANCE
BANDWIDTH (GHz)	2.5	2.5
MAX POWER OUTPUT (W)	0.5	0.25
EFFICIENCY AT MAX OUTPUT (%)	15	< 10
GAIN (dB)	20	20
4TH STAGE GATE WIDTH (MM)	----	1.2

High Power 20 GHz Amplifier

A second 20 GHz chip focussed entirely on the production of the maximum possible 20 GHz power in a monolithic chip. Once more the contractor was Texas Instruments. This effort resulted in a three stage amplifier using single gate FETs. In order to achieve the desired power output, each stage features multiple parallel gates (as many as 32 for the output stage). The chip performance is summarized in Table 3.

TABLE 3.
TEXAS INSTRUMENTS 20 GHZ HIGH POWER AMPLIFIER

	DESIGN GOAL	PERFORMANCE
BANDWIDTH	2.5	2.5
MAX POWER OUTPUT (W)	2.5	> 2.0
EFFICIENCY AT MAX POWER (%)	20	16
GAIN (dB)	15	18
LAST STAGE GATE WIDTH (MM)	----	3.6

20 GHz Integrated Transmitter Module

Ultimately, one desires to incorporate all the elements of a transmitter module (phase shifter, variable power amplifier, and power amplifier) on a single chip. The advantages of such high level integration include: improved reliability, compactness, potential performance enhancements, and reduced cost. Such a project was pursued under contract with Rockwell. The chip produced a 21 dBm output power with 15 dB gain. A picture of the module developed under this effort is shown in Figure 2.

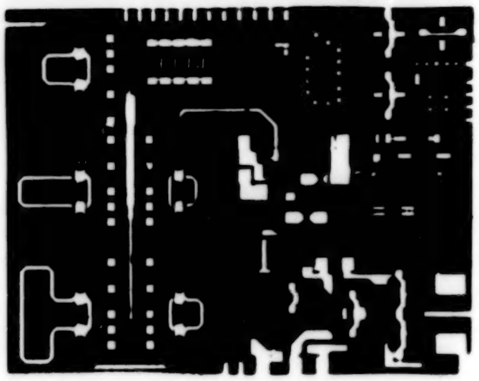


FIGURE 2. ROCKWELL 20 GHz TRANSMIT MODULE

30 GHz Integrated Receiver

Active phased arrays have been proposed for a number of NASA applications, including: array feeds for deep space communications, multiple-beam satellite communications, hemispherical coverage multiple access communications for Space Station Freedom, orbital debris tracking for Space Station Freedom, and adaptive arrays for distorted reflector compensation. To enable the practical implementation of scanning arrays, the complexities of device and antenna integration must be solved. As a preliminary step, prototype 30 GHz receiver modules were developed by Hughes and Honeywell. Honeywell developed an interconnected MMIC receiver for which all functions were performed at the RF frequency. The device consisted of a two-stage 0.25 by 100 micron gate FET low noise amplifier (LNA), a two-stage dual-gate gain control amplifier, and a four bit switched-line/loaded-line phase shifter. The receiver produced a noise figure of 14 dB at maximum gain (13 dB) and achieved full 360 degree phase coverage in nominally 11.25 degree increments. Ultimately, the LNA would require six stages of amplification, which would reduce the noise figure, in principle, to about 5 dB. Hughes was directed to implement phase shift control at the LO frequency and gain control at the IF. The advantage of this approach is that the array beam forming network can be implemented at lower frequencies, albeit the benefit could be offset by the need for a mixer at each antenna element. In fact, the mixer was perhaps the most troublesome component in both designs, providing a best case conversion loss of 8 dB.

HIGH PERFORMANCE MONOLITHIC CIRCUITS

Background

It is clear that all of the GaAs MESFET-based modules, although they constituted benchmark achievements at these frequencies, suffered from many of the same problems as the earlier hybrid implementations of solid state technology. For the power devices, their efficiencies make their use marginal for space applications, except in very limited numbers, such as would be required if they were used as a driver for a higher efficiency final stage. An attempt to use multiple chips with any kind of combining would lead to prohibitively large prime power requirements. For receiver modules, the noise figures obtained using MESFET technology are not competitive with that which can be attained using discrete devices and custom-tuned circuits. However, recent advances in semiconductor materials, enabled by the development of molecular beam epitaxy (MBE) techniques, have drastically improved the performance which basic devices can achieve. Typically, power-added efficiencies near 50% can be obtained for a single device at frequencies near 30 GHz. While device noise figures less than 1 dB are possible at 60 GHz.

At the present time NASA Lewis is sponsoring the development of monolithic chips based on heterojunction devices. Two of these at 32 GHz are for possible use in the space (transmitter) portion of the deep space communications network. The third and fourth at 60 GHz are intended for application to intersatellite communications, such as might be required by Advanced TDRSS or lunar/Mars exploration. A fifth chip will operate at 95 GHz, with potential applications in interplanetary communications, or in earth-observation systems.

32 GHz Amplifiers

In another application the NASA deep space communications network is considering a conversion to Ku-band. The primary motivation for such a change is the significant increase in antenna gain (for a fixed aperture size) and the corresponding decrease in power requirements (for a fixed data rate). Increased antenna gain, however, implies smaller beams and therefore more stringent pointing requirements. Such a situation, of course, is ideal for implementation of an electronically steerable phased array, which does not disturb other critically-pointed spacecraft instruments (experiments or sensors) in the way a mechanically steered antenna would. To support breadboard evaluations of such a system, 32 GHz power amplifier modules are under development. The contractors executing these efforts are Texas Instruments and Hughes Aircraft. The TI work has been under way since May, 1985, and is near completion, while the Hughes effort was initiated in June 1988.

TI proposed and originally designed amplifiers using GaAs MESFET technology, but was directed, after approximately 18 months work, to concentrate on heterojunction devices. Specifically, they have investigated AlGaAs/GaAs HEMT structures and pseudomorphic InGaAs/GaAs structures. At this point it is clear that the pseudomorphic technology outperforms both the AlGaAs HEMT and the GaAs MESFET technology by a significant margin at 32 GHz. The specific pseudomorphic structure which TI has adopted is shown in Figure 3.

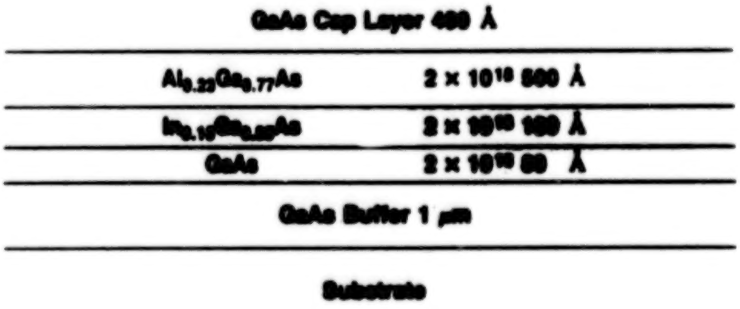


FIGURE 3. TEXAS INSTRUMENTS' PSEUDOMORPHIC POWER AMPLIFIER STRUCTURE

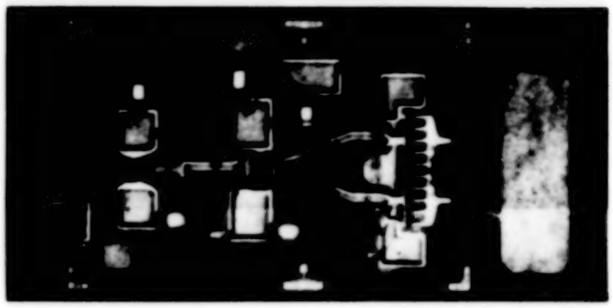
ORIGINAL PAGE IS
OF POOR QUALITY

The performance parameters for two of the pseudomorphic chips developed under this program are shown in Table 5.

TABLE 5.
TEXAS INSTRUMENTS 32 GHz MONOLITHIC POWER AMPLIFIER PERFORMANCE

	3-STAGE AMP	1-STAGE AMP
BANDWIDTH (GHz)	2.0	2.0
GAIN (dB)	23	4.6
GATE LENGTH (uM)	0.25	0.25
FINAL STAGE GATE WIDTH (MM)	.25	.25
POWER OUTPUT (mW)	190	460
EFFICIENCY (%)	30	24

The layout of the three-stage 2.6 mm by 1.2 mm MMIC amplifier is shown in figure 4.



ORIGINAL PAGE IS OF POOR QUALITY

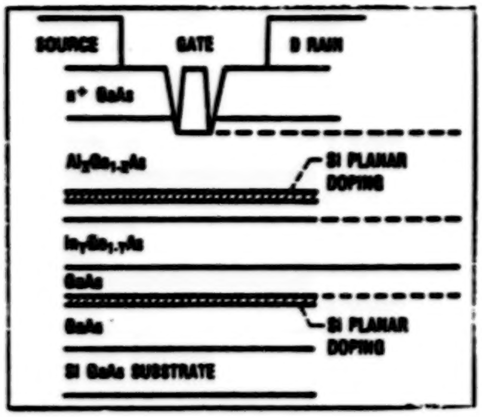
FIGURE 4. TI's 3-STAGE MONOLITHIC 32 GHz AMPLIFIER

In a parallel 32 GHz effort Hughes Aircraft Corporation's Microwave Products Division and Malibu Research Laboratories are collaborating on the development of a 32 GHz variable power amplifier. The design goals for this chip are shown in Table 6.

TABLE 6.
DESIGN GOALS FOR HUGHES 32 GHz VARIABLE POWER AMPLIFIER

BANDWIDTH (GHz)	2.0
MAX POWER OUTPUT (mW)	150
EFFICIENCY AT MAX POWER(%)	40
GAIN AT 1 DB COMP.(dB)	15.0

The Hughes contract, like most such developments at this time is to be carried out in several phases. These will consist of (1) the optimization of a single gate device design; (2) the development of a single stage amplifier; (3) the development of a dual gate device; (4) the design, fabrication, and test of a three-stage, single gate amplifier; and finally, (5) a three-stage dual gate amplifier. In the 16 months that the Hughes team has been under contract, they have carried out the first two phases. The epitaxy which they have selected for the basic device is similar to that utilized by TI, except that Hughes has elected not to dope the active layer. It does, however, utilize a single active InGaAs layer with donors on each side. The structure and performance parameters for the basic 32 GHz device are shown in Figure 5.



FREQUENCY	32.0
GAIN*	4(5) dB
POWER OUTPUT*	222(123)mW
EFFICIENCY*	23(41)%
GATE WIDTH	300 uM
GATE LENGTH	0.2 uM

*tuned for max power(eff)

FIGURE 5. STRUCTURE AND PERFORMANCE OF HUGHES 32 GHz POWER MODFET

This device has been incorporated into a single stage amplifier which exhibited an output power of 125 mW at 21% efficiency with 5.5 dB gain. This amplifier is intended as the third stage of the final monolithic module. These results represent the first iteration of this chip, and significant improvement is expected before the program ends.

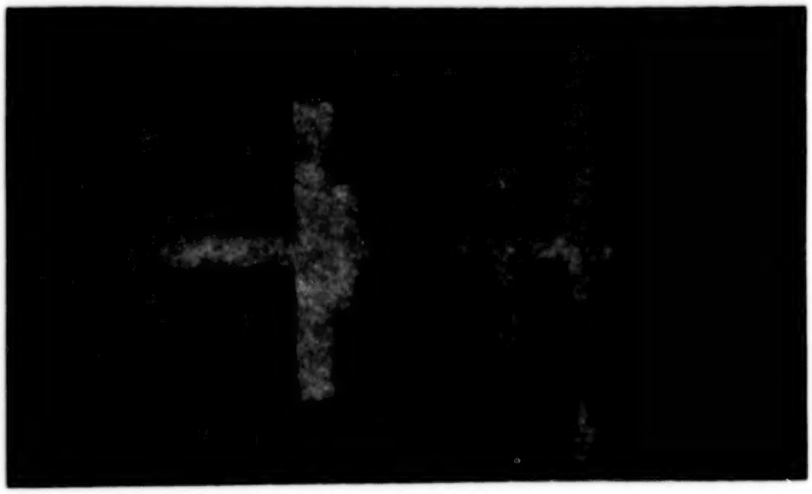
60 GHz Amplifiers

In addition, under the same contract, Hughes is developing a 60 GHz monolithic power amplifier. The justification for this program is eventual application in intersatellite links. Although NASA's plans for the Advanced Tracking and Data Relay Satellite (ATDRS) do not presently call for 60 GHz crosslinks, it seems likely that if such technology were available it would eventually find application in that area. The performance goals for the program, are shown below in Table 7.

TABLE 7.
Performance Goals for Hughes' Pseudomorphic 60 GHz

	Power Amplifier
BANDWIDTH (GHz)	2.0
MAX POWER OUTPUT (W)	0.1
GAIN (dB)	15
EFFICIENCY (%)	30.0

At 60 GHz Hughes is using the same basic pseudomorphic device structure as at 32 GHz, although the gate lengths have been shortened somewhat (0.1 to 0.15 uM). The layout and the performance achieved for a single stage amplifier are shown in Figure 6. The chip is approximately 1.5 mm long. As in the 32 GHz module, the amplifier shown is intended as the third (high power) stage of the completed monolithic amplifier.



POWER OUT	112 mW
GAIN	6 dB
EFFICIENCY	26%

FIGURE 6. HUGHES SINGLE STAGE MONOLITHIC 60 GHz AMPLIFIER

ORIGINAL PAGE IS
OF POOR QUALITY

FUTURE ACTIVITIES

Phased Array Development

The 32 GHz power amplifier modules developed by Texas Instruments and described here are scheduled to be incorporated into a breadboard transmitter array antenna which will also utilize phase shifters developed under NASA Lewis sponsorship. This work is being carried out at NASA's Jet Propulsion Laboratory, where a two-dimensional array is expected to be completed late in calendar year 1990.

The Hughes work at both 32 Ghz and 60 Ghz is probably at least a year away from being used even in a breadboard system. Although the contract is scheduled for completion in early 1991, it has yet to address what has been one of the major difficulties in the fabrication of a multistage power amplifier - inadequate large signal device models. It has been a common experience for a designer to develop excellent individual stage amplifiers, which meet all the requirements of the overall power and gain budget, only to find that the multistage module performance falls far short of the program requirements. Consequently, it appears optimistic to expect that Hughes will complete their development by 1991. 1992 would appear to be more realistic. At that time, it is anticipated that a 60 GHz breadboard array will be built, either at JPL or at NASA Lewis. As with the Ku-band array, it will utilize monolithic phase shifters which are being developed in parallel at Hughes.

W-Band Receiver

The most recent MMIC developmental effort is the production of a 94 GHz receiver. Potential applications include deep-space communications, radiometry, and orbital debris tracking radar. Goals of the program include a noise figure of 3.5 dB and a gain of 18 dB. It is anticipated that the ambitious goals of this program will necessitate a technology based on InP, rather than traditional GaAs.

REFERENCES

1. Saunier, P.; Tserng, H. Q.; and Kim, B.: Monolithic Gallium Arsenide Variable Power Amplifier. NASA CR 180814, 1987.

2. Higgins, J. A.: 20 GHz Monolithic Transmit Modules. NASA CR 182134, 1987.

3. Geddes, J.; Mondal, J.; Contolatis, T.; and Sokolov, V.: 30 GHz Monolithic Receive Module. NASA CR 185198, 1989.

4. Liu, L.; Lin, C.; Kessler, J.; and Wang, S.: A 30 GHz Monolithic Receiver. IEEE Microwave and Millimeter-Wave Monolithic Circuit Symposium Digest, 1986, pp. 41-44.

N91

24070

UNCLAS

N91-24070

OPTICAL COMMUNICATION FOR SPACE MISSIONS

M. Fitzmaurice
NASA/Goddard Space Flight Center
Greenbelt, MD 20771

Abstract

NASA/GSFC has played a leading role in the development of direct detection (DD) optical communications for space applications. The key challenges for optical communications have been in the development of reliable optical power sources and in the design of high performance pointing, acquisition and tracking systems mandated by the narrow widths of transmitted optical beams. These areas have been the focus for GSFC involvement in optical communications over the past few years. GaAlAs diodes and diode arrays are the most attractive technology for optical transmitters, and accordingly, GSFC has conducted extensive diode life testing and performance characterization studies on commercially available laser diodes. Pioneering work on dichroic and grating techniques for combining the power of several laser diodes has also been carried out. Other technology development activities at GSFC have included work on injection locked diode arrays for far-field pattern improvement, novel implementation of laser diode modulators, and development of advanced optical detectors for communications and tracking. GSFC work on optical systems development has been centered around brassboard hardware which has been developed to demonstrate optical link acquisition and submicrodian tracking/pointing performance. This system is referred to as the Pointing, Acquisition and Tracking System (PATS). A detailed computer simulation model of the PATS has been developed, and an extensive program has been initiated that will compare test and simulation data, and will provide an in depth understanding of the system design. Other systems work supported by GSFC has included the development of a 50 Mbps optical communications system demonstrating a bit error probability of 10^{-6} with a signal level of 55 photo-electrons per bit. Computer models have also been developed that describe the average and burst error statistics of the optical communications channel. The overall GSFC program supporting optical communications technology and systems is directed at a flight demonstration in the mid to late 90's. Space Station Freedom will serve as the space platform for this experimental flight system.

1. Overview

Optical communications systems offer the benefits of much smaller apertures and reduced weight and power in comparison with microwave systems, particularly at data rates in excess of 100 Mbps. Optical systems also avoid problems of RF spectrum allocation and interference. Accordingly, optical technology is a prime candidate for satisfying emerging very high data rate relay needs of NASA and the commercial satellite industry.[1,2] Direct detection (DD) optical systems also offer the benefits of simplicity relative to heterodyne systems. DD optical power need not be coherent and this enables the applications of power summing techniques which have demonstrated 200 mW beam powers and promise more than 1 W in the next few years. Furthermore, DD requirements for optical frequency stability are not as strict as those for heterodyne systems resulting in relatively wide tolerances on the choice of laser sources (e.g. combined diodes or multispectral mode) and on the temperature control of operating diodes. Finally, a DD receiver is significantly less complex than heterodyne, in that it does not track the optical frequency and need not be designed to diffraction limited tolerance.

The overall GSFC program in direct detection optical communications is aimed at developing the technology to meet the needs of NASA and the communications satellite industry. In Section 2, the GSFC view of the resulting requirements for optical systems at the component level and the system level is discussed. Providing assurance, via component technology development and system demonstration, that these requirements can be satisfied has been the major thrust of the GSFC program in optical communications. GSFC activities have included the development and testing of key components (discussed in Section 3) as well as the implementation of systems to demonstrate critical end-to-end performance requirements (discussed in Section 4). Increasingly, the primary goal of the program has become the development of proof of concept optical communications systems that will by form, fit, and function, demonstrate the readiness of the technology to support NASA requirements and those of the satellite communications industry. For the near term, such activities will be limited to ground based demonstrations due to resource limitations. However, the overall program is geared toward a flight demonstration of an experimental system as an Attached Payload on Space Station Freedom (SSF). This flight experiment is described in Section 5.

2. Requirements for Optical Communications Systems

2.1 Data Rates and Channel Quality

The primary requirements of NASA and commercial missions addressed by optical communications systems is the need for high data rate (100's of Mbps to Gbps) ISLs from LEO to GEO or across the qeosynchronous arc. The required channel quality varies from 10^{-6} to 10^{-9} BER depending upon the specific mission. Optical systems are attractive for this application because, as indicated in Figure 1, the required telescope apertures are quite small relative to the needed antenna apertures for RF systems. Reliable optical power is clearly one of the key requirements for optical communications, and GaAlAs laser diodes in the 0.8 to 0.9 um region are the prime candidates for this function. Performance requirements include high output powers in a single spectral and spatial mode, and stability under high data rate modulation with rise and fall times in the order of 400 ps. Furthermore, long lifetimes are required in order to support 5 to 10 year missions. With currently available single laser diodes and SI APD detectors, 30 cm to 50 cm apertures are required to relay 100 Mpbs to 1 Gbps across the LEO-GEO range. Power combined laser diode transmitters reduce the requirement to between 20 cm and 30 cm. Next generation improvements in laser diodes and detectors will permit even smaller apertures and a corresponding relaxation of pointing performance (Figure 2). Advances in single laser diodes, diode arrays, and power combining techniques, promise powers above 1W. In addition, reliable high peak power operation laser diodes will be able to reap the full advantages of quaternary pulse position modulation (PPM) over binary PPM. Finally, greater than 3 dB detector sensitivity improvement is likely from a new class of photodetectors currently under development.[3]

2.2 Spatial Acquisition

The small wavelength of optical communications systems result in extremely narrow beams which require precise pointing at the submicroradian level. Such accuracy is not obtained via open loop pointing so that establishing any optical communications link involves a spatial acquisition scenario in which the initial 1-10 mrad open loop pointing uncertainty typical of space platforms is reduced to <1 urad via mutual closed loop tracking. In an acquisition scenario, two cooperating terminals orchestrate a mutual scan/search. Typical scenarios contain multiple stages where initial broad beacons are successively narrowed to the diffraction limit as pointing accuracy is improved via closed-loop tracking. Achieving spatial acquisition reliably and within a short time (< 30 sec) is a strong driver for both low noise photosensitive area arrays, and for high power laser sources which can provide adequate signal strengths at GEO distances even broadened to 1 mrad and beyond.

2.3 Fine Tracking and Pointing

Typical space platforms contain high frequency vibrations which will create significant pointing jitter for attached instruments. For example, measurements on LANDSAT identified a pointing jitter of approximately 11 urad integrated out to 125 Hz. Clearly, if submicroradian pointing accuracy is to be achieved, such jitter must be suppressed by some 10 to 20 dB via pointing servomechanisms. Although there are many alternative approaches, the most commonly suggested one involves high bandwidth spatial tracking of the received beacon. With the transmitted and received signal sharing a common path the transmitted signal pointing is automatically compensated by maintaining the received signal spot at a reference position. The need to achieve 20 dB of jitter rejection out to a few hundred Hz is a driver for high performance pointing systems with servo bandwidths of 1000 Hz or more. In addition, high sensitivity tracking detectors with bandwidths of 5,000 to 10,000 Hz are needed.

2.4 Transmit Point-Ahead Correction

Because of the narrow beams and the fact that the platforms are in relative motion, a transmit pointing bias relative to the direction of the received signal is required. The required bias is equal to 2 V/c, where V is the relative velocity of the platforms perpendicular to their line of sight and c is the speed of light. For typical LEO to GEO applications, this tends to be on the order of ± 80 urad. Implementations for pointing-ahead ted to complicate the focal plane of the optical transceiver. One GSFC activity is exploring a simplified point-ahead scheme using a tracking area array detector.

3. Component Testing and Development

3.1 Laser Diodes Performance Characterization and Life Testing

Because the optical power source is perhaps the most critical component of an optical communications systems, GSFC began in 1984 a comprehensive testing program for single laser diodes.[4,5] The goal of this program has been to procure and evaluate commercially available high power GaAlAs laser diodes. For the Sharp (LT015MD) and Hitachi (HL8314E) laser evaluated at the GSFC, typical output powers are 30 mW average, 50-60 mW peak under 100 MHz square weave modulation, 50 percent duty cycle. These lasers are mostly single spectral mode, emitting between 810 and 840 nm, and single spatial mode, emitting in far-field patterns typically 10° by 25° at FWHM. Their polarization ratios are between 100:1 and 200:1. Wavefront errors are approximately lamda/30. Typical efficiencies are 0.7-0.9 W/A for sharp lasers and 0.4-0.6 for Hitachi. Results of these tests are summarized in Tables 1 and 2.

3.2 Laser Diode Power Combining

In order to achieve optical transmitter powers of several hundred milliwatts, GSFC has examined two combining techniques which incoherently sum the outputs of single element laser diodes. The first technique uses diffraction gratings to combine diodes in parallel while the second technique uses dichroic filters to combine diodes in series. Both techniques require stable wavelength separation of the individual laser diodes for efficient combining. Incoherent summing of single element laser diodes also builds in redundancy should one or two diodes fail catastrophically.

The Grating Laser Beam Combiner (GLBC) design uses the first-order diffraction from each grating to combine the laser outputs. This limits the overall efficiency to no better than the product of the first-order diffraction efficiencies of each grating. Current grating technology is capable of first-order diffraction efficiencies on the order of 85%. This limits the GLBC throughput efficiency to about 70%. These and other aspects of the GLBC are described in detail elsewhere.[6] A four laser diode GLBC has been built at GSFC and is nearing the test and evaluation stage. Figure 3 is a schematic of this GLBC.

An alternative to the GLBC is a combiner which employees dichroic filters to coalign the collimated outputs of multiple laser diodes. A proof of concept (POC) prototype using this approach has been developed by McDonnell Douglas Astronautics Company under contract to GSFC.[7] The Laser Power Summing System (LPSS) employs dichroic filters to spatially combine the output from seven laser diodes spaced at 2 nm intervals. Since this is a serial combining technique, the losses cascade such that the n^{th} laser diode reflects off n-1 dichroic filters. The losses multiply and soon a point of diminishing returns is reached such that the small increase in power afforded by the n+1 diode is offset by the added weight and optical alignment complexity. The practical upper limit appears to be between 10 and 15 laser diodes.

The optical system layout of the LPSS is depicted schematically in Figure 4. As shown, seven laser diodes are arranged along successive reflection lines of the dichroic filters. The performance of the final prototype system is summarized in Table 3. As indicated, the LPSS produced in excess of 170 mW of optical power from seven single element laser diodes with a composite beam divergence within 20% of the diffraction limit of a single channel.

4. System Development and Simulations

4.1 The GSFC Pointing, Acquisition and Tracking System (PATS)

The goal of the PATS program is to create a testbed to support design and specification of space-based optical direct detection communications terminals, and to demonstrate the

required spatial acquisition and tracking performance. The PATS is a brassboard/breadboard configuration of hardware capable of emulating space-based optical terminal pointing, acquisition, and tracking, and is supported by a detailed computer simulation of the hardware performance.[8]

4.1.1 Hardware Description

The PATS hardware currently consists of two terminals; a beacon simulator and a transceiver simulator, as depicted in Figure 5. The beacon simulator consists of a laser, two galvanometers which drive fine pointing mirrors, a noise generator, and a quad cell, and various optics. The beacon simulator initiates a test scenario by sending a laser beam to the transceiver. Two mirrors mounted on galvanometers induce spatial jitter in azimuth and elevation to simulate transceiver platform attitude motion. The transceiver locates and tracks the beacon simulator laser. During this process, the transceiver will send back a beam to the beacon simulator so that the transceiver nested control loop error can be measured.

The beacon quad cell senses the return beam and produces an error voltage proportional to the transceiver off-point angle. Either an oscilloscope or a monitor can be used to display this error signal. The transceiver simulator consists of the following components: a laser, two galvanometers, a vernier pointing system (VPS) controller, a gimbal assembly, a transconductance stage, a digital encoder, an all digital controller (ADC), a computer, a quad cell detector, a charge coupled device (CCD) camera, various optics, and several electronic modules. Figure 6 shows a view of the transceiver.

4.1.2 Preliminary Performance Analysis

The compensation for high frequency jitter is one of the key performance requirements for optical communications systems that PATS will demonstrate. For example, pointing jitter induced by space platform vibration can exceed 10 urad. Such jitter will be suppressed via high bandwidth tracking of an incoming beacon that will compensate for pointing jitter. Figure 7 is a plot of the disturbance rejection spectrum $|1-H(s)|^2$ for the current PATS configuration where H(s) is closed loop frequency response of the PATS pointing servomechanism. The 300 Hz resonance of the transceiver fine steering galvanometer is clearly visible. Figure 8 is a computer simulation output modeling PATS which shows how it would compensate for the pointing jitter power spectral density (PSD) measured on-board LANDSAT. The first frame of Figure 8 is the LANDSAT PSD which represents more than 11 urad rms jitter out to 125 Hz. The second frame shows the relevant portion of the PATS disturbance rejection spectrum. The third and final frame shows the PSD of the pointing error or untraced platform jitter. The total rms pointing jitter is roughly 0.5 urad and is dominated by the highest frequency components. It is interesting to note that the 0.5 u rad value is roughly the requirement for an optical system with a 20 cm aperture.

4.1.3 High Data Rate Transmitter/Receiver Systems

GSFC is supporting the development and testing of direct detection breadboarding systems using commercially available components. The prototype systems use a GaAlAs laser diode transmitter (lamda=834 nm) of 30mW maximum average power and a silicon avalanche photodiode (APD) GaAsFET feedback resistor type preamplifier. The signal format is a 4-slot pulse-position modulation (PPM), in which each pair of source data bits is encoded as an optical pulse in one of four slots within each transmitted word. A 50 Mbps system has been developed and has demonstrated a performance only 9 dB above the quantum limit by achieving a bit error rate (BER) of 10^{-6} at a received signal level of 55 detected photons per bit.[9]

Figure 9 illustrates the overall receiver design.[14] The optical signal is detected, amplified and split among sections which perform the maximum likelihood (ML) data estimation, slot synchronization, and word synchronization. For rectangularly shaped optical pulses, the ML estimator integrates the APD/preamp output signal over each time slot and then determines in which slot the largest value occurred for each word. For a source data rate R bits/sec and 4-PPM signal format, the slot duration $T_s = 1/(2R)$. The matched filter approximates an ideal integrator of duration T_s by splitting the input signal into four cable which have relative delays of 0, $T_s/4$, $T_s/2$, and $3T_s/4$, and then passing the sum of the delayed signals through a low pass filter. The matched filter output is then split among four cables having relative delays of 0, Ts, $2T_s$, and $3T_s$ whose outputs are compared at the proper instant to give an estimate of the received PPM word. The APD/preamp output is also sent through a pulse shaper which drives a phase lock loop (PLL) in order to recover the slot clock. The PPM word clock is recovered by locking a second PPL to the occurrence of back-to-back PPM pulses. No degradation in BER is evident with the recovered synchronization as compared to perfect (hard wired) synchronization when the received optical power is sufficient to produce 10^{-6} BER.

Theoretical models of the receiver were developed which accurately predict optimal system parameters and BER performance, A non-Gaussian model was necessary to simulate the APD output photocurrent. Use of the Gaussian model under conditions of negligible background radiation (less than one detected noise photon per slot) and very low APD dark current lead to an underestimate of the optimal APD gain, and therefore the BER would be higher than what is actually achievable when the proper gain is used. Figure 10 compares measured data with theoretical calculations of BER.

5.0 Proposed Flight Demonstration of Optical Communications

GSFC has proposed the development of laser communications transceiver (LCT) for Space Station Freedom which will be used to conduct a broad class of experiments in optical communications. The major areas of inquiry will include the achievable performance if a space-based LCT with regard to rapid beacon signal acquisition, fine tracking and pointing with submicroradian accuracy, and wideband communications at Gbps data rates with bit error rates as low as 10^{-9}. This will demonstrate the application of the LCT technology to future commercial and NASA communications needs including high data rate GEO-GEO and LEO-GEO communications links.

The approach toward the LCT design is to build a single optical bench and associated optics, including a 20 cm gimbaled telescope, which may be shared by alternative transmitter and receiver modules that embody competing technologies. Candidates for the transmitter modules include single GaAlAs diodes, multiple diodes power combined (for increased power) via dichroic beam splitters or gratings, and monolithic arrays of diodes. Candidates for tracking receiver modules include quadrant silicon avalanche photodiodes and high bandwidth charge coupled arrays.

A range of experiments will be conducted with the LCT in a variety of operational modes. The most important operational mode of the LCT is in concert with another cooperating laser transceiver where the overall performance of a mutual acquisition sequence, mutual tracking/pointing, and duplex communications can de demonstrated and measured. In fulfillment of these goals, GSFC is committed to upgrading its ground-based optical site to perform compatible LCT functions. Furthermore, GSFC is seeking additional opportunities for another laser transceiver in space at LEO (e.g., as a Shuttle attached payload) or at GEO (e.g., on NASA, ESA, or Japanese data relay satellite). A prime GEO opportunity is offered by the Advanced tracking and Data Relay Satellite (ATDRS) which will support a Pre-Operation Demonstration (POD) of a novel communications concept. Two of the three POD options under consideration involve laser communications: one would demonstrate a 650 Mbps LEO-ATDRS optical link, and the other a 2 Gbps ATDRS-ATDRS crosslink.

REFERENCES

1. Intersatellite Link Application to Commercial Communications Satellites, NASA CR-179-598, prepared for NASA/LeRC by COMSAT, January 1987.

2. The ATDRSS Mission Model: Projected User Requirements for 1996-2010, prepared for NASA/GSFC by Stanford Telecommunications, Inc., January 1988.

3. F. Capasso, W. Tsong and G. Williams, "Staircase Solid State Photomultipliers and Avalanche Photodiodes with Enhanced Ionization Ratios," IEEE Trans. Electron Dev., Vol. ED-30, No. 4, pp. 381-390, April 1983.

4. K. A. Forest, J.B. Abshire, S.S. Manizade and P.L. Spadin, "Characterization and Detailed Lifetesting of High Power 830 nm AlGaAs Laser Diodes," SPIE Symposium on Innovative Science and Technology for Government and Civilian Applications, January 1989.

5. T.L. Holcomb and K. Forrest, "Space-Qualified Laser Diode Assembly for the NASA/DDLT Program", Proc. SPIE, January 1989.

6. P.O. Minott and J.B. Abshire, "Grating Rhomb Diode Laser Power Combiner," SPIE Vol. 756, Optical Technologies for Space Communications, January 1987.

7. Proof of Concept Model Semiconductor Laser Power Summing System, prepared for NASA/GSFC for McDonnell Douglas Astronautics Co., June 1986.

8. N.D. Fox, W.W. Chapman and T.S. Johnson, "Performance and Modeling of the GSFC Pointing, Acquisition and Tracking System for Laser Communications," Proc. SPIE, January 1989.

9. X. Sun, F. Davidson, and J.B. Abshire, "Performance Measurements of a Diode Laser Optical Communications Link with Q=4 PPM Signaling, Optical Technologies for Space Communications Systems, ed. K. Bhasin and G.A Koepf, pp. 50-53, Proc. SPIE 756, 1987.

TABLE 1: FAR-FIELD PATTERN DRIFT

FAR-FIELD PARAMETER		HITACHI HL8314E	SHARP LT015MD
$\frac{d(\text{mean})}{dt}$ (°/KHr)	0°	0.021 ± 0.015	0.007 ± 0.007
	90°	0.028 ± 0.031	0.009 ± 0.009
$\frac{d(\sigma\theta)}{dt}$ (°/KHr)	0°	0.051 ± 0.070	0.083 ± 0.084
	90°	0.017 ± 0.039	0.031 ± 0.024
AVERAGE FAR-FIELD CHANGES: PEAK-WANDERING AND CHANGES IN FWHM:			

TABLE 2: DIODE SPECTRAL STABILITY

DIODE	HOURS ($\times 10^3$)	MODAL DRIFT (nm/KHr)	MODE HOPS (NO. AND WAVELENGTH)
HITACHI			
7C0766	2.8	0.011 ± 0.028	(3) 834.7 -> 834.4 -> 834.7 -> 834.4
7C0765	4.5	0.031 ± 0.051	(many) 832.84 min -> 833.0 max
7C0764	4.5	0.0035 ± 0.028	(13) 833.2 min -> 835.0 max
7C0763	4.5	0.05 ± 0.043	(7) 833.2 min -> 833.7 max
7B1568	5.8	0.01 ± 0.034	(18) 840.8 min -> 841.7 max
		Avg. = 0.02113 Stand. Dev. = 0.034	
SHARP			
76-79310	4.5	0.0097 ± 0.033	(7-8) 834.8 min -> 835.0 max
70-79310	4.5	− 0.035 ± 0.036	(5) 832.8 min -> 834.2 max
69-79310	4.5	0.0082 ± 0.003	(2) 830.1 min -> 832.6 max
56-79310	6.2	0.0054 ± 0.027	(2) 834.3 min -> 834.9 max
52-79310	2.1	0.00071 ± 0.021	(2) 834.5 min -> 836.2 max
162-74530	4.0	- 0.0026 ± 0.020	(12) 828.6 min -> 830.6 max
		Avg. = 0.0118 Stand. Dev. = 0.0134	

TABLE 3
LASER POWER SUMMING SYSTEM PERFORMANCE CHARACTERISTICS

10−12937

ITEM	WORK STATEMENT	LPSS
NO. OF DIODES	7	7
WAVELENGTH	$0.8 < \lambda < 0.9$	.83 ± .005
DIODE POWER	$P_o < 20$mW*	$P_o > 30$mW
OPTICAL THRUPUT	>80	> 80%
MODULATION BANDWIDTH	250 MHz	52 kHz – 800 MHz
OUTPUT BEAM DIAM	0.8< d <1.2 cm	1.2 cm
CO-LOCATION OF BEAMS	±5% OF BEAM DIAM	± .05 cm
OUTPUT BEAM DIVERGENCE	NOMINALLY DIFFRACTION LIMITED	< 1.2 TIMES DIFFRACTION LIMIT
CO-ANGULARITY OF COMBINED BEAMS AT OUTPUT WINDOW	±5% OF BEAM DIVERGENCE	100 MICRORADIANS
EXTINCTION RATIO	—	> 20:1
MASS	5 kg	4.8 kg
VOLUME	.0125 m³	.0062 m³
OPERATING TEMP.	AMBIENT	AMBIENT
WARMUP TIME	t_w = 30 MIN	t_w = 8 MIN
UNATTENDED OPERATION	t_o > 60 MIN	t_o > 24 HRS

ORIGINAL PAGE IS
OF POOR QUALITY

CURRENT TECHNOLOGY LINK ASSESSMENT
(40,000 km link; 10e-6; 6 db Margin; k=0.01; BPPM)

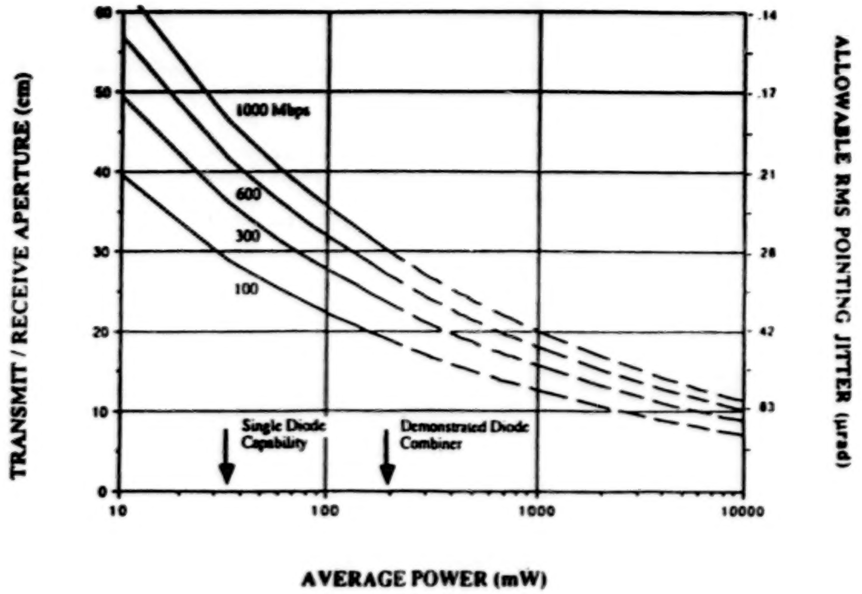


FIGURE 1

NEXT GENERATION LINK ASSESSMENT
(40,000 km link; 10e-6 BER; 6 db Margin; F=1; QPPM)

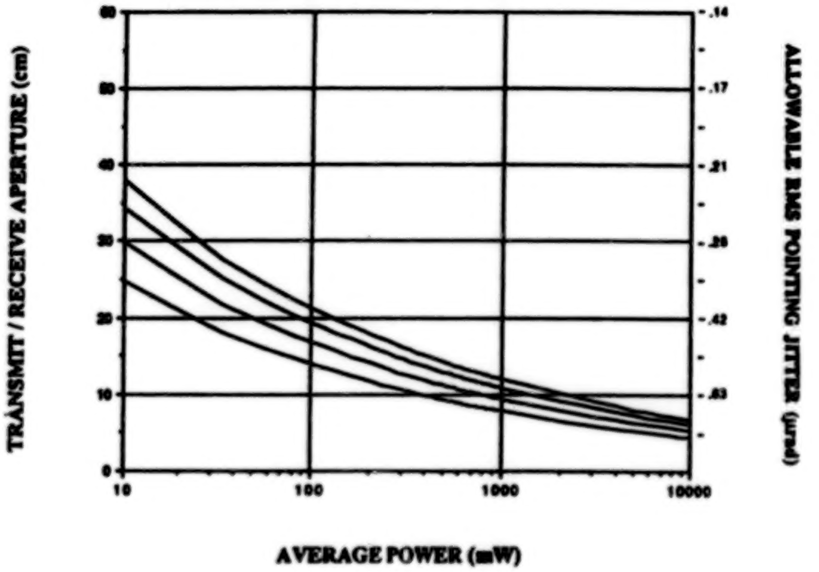


FIGURE 2

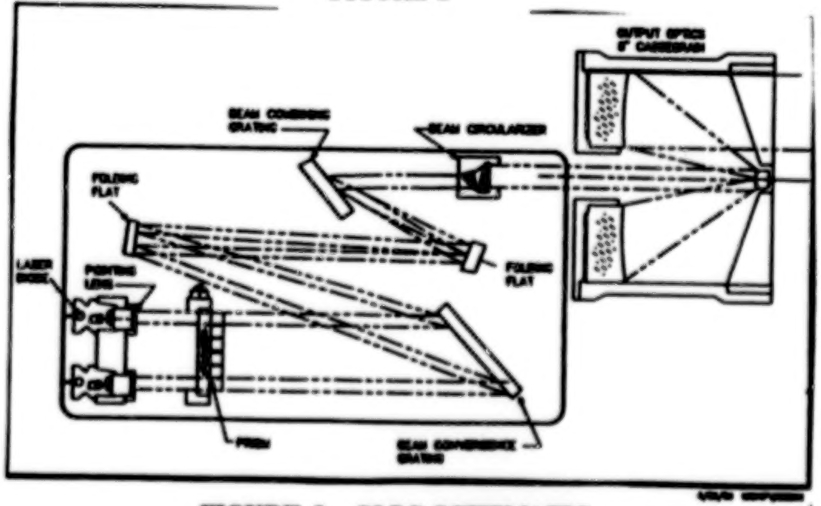

OUTPUT OPTICS 8" CASSEGRAIN
BEAM COMBINING GRATING
BEAM CIRCULARIZER
FOLDING FLAT
LASER DIODE
POINTING LENS
FOLDING FLAT
PRISM
BEAM CONVERGENCE GRATING

FIGURE 3: GLBC SCHEMATIC

ORIGINAL PAGE IS
OF POOR QUALITY

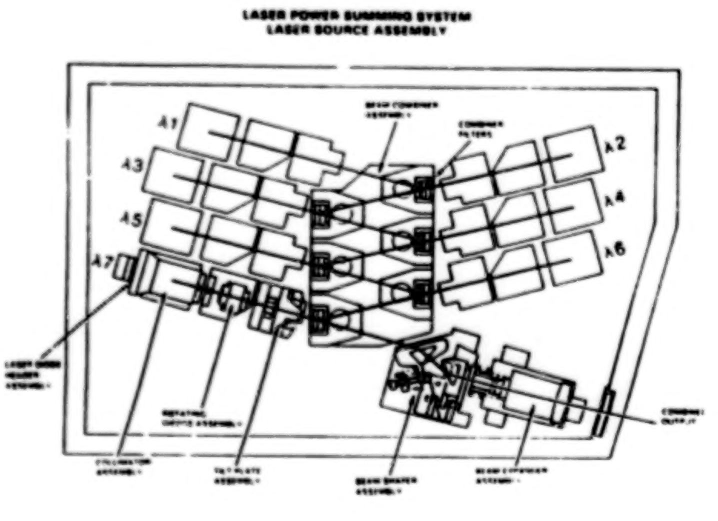

LASER POWER SUMMING SYSTEM
LASER SOURCE ASSEMBLY
A1
A2
A3
A4
A5
A6
A7
BEAM COMBINER ASSEMBLY
COMBINER FILTERS
LASER DIODE HEADER ASSEMBLY
ROTATING OPTICS ASSEMBLY
COLLIMATOR ASSEMBLY
TILT PLATE ASSEMBLY
BEAM SHAPER ASSEMBLY
BEAM EXPANDER ASSEMBLY
COMBINED OUTPUT

FIGURE 4

PATS Transceiver

FIGURE 5

PATS Hardware Configuration

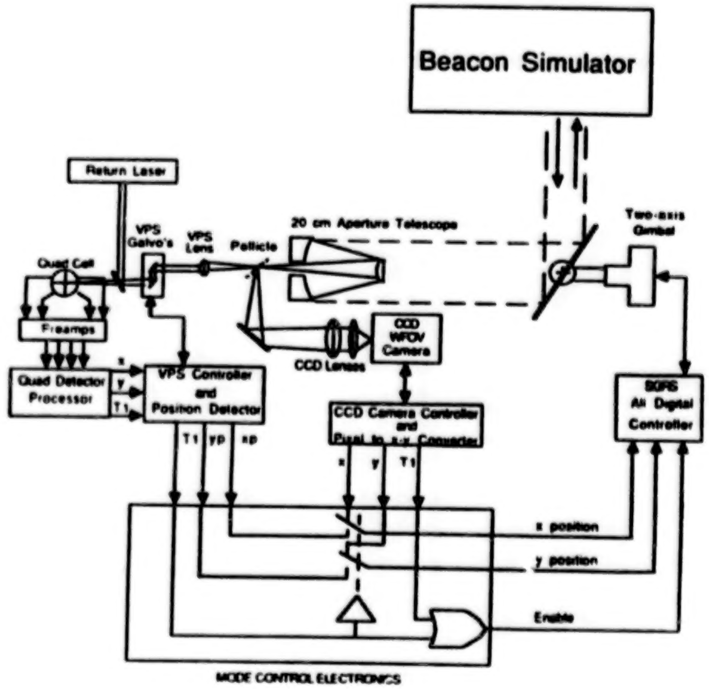

Beacon Simulator
Return Laser
Quad Cell
VPS Galvo's
VPS Lens
Pellicle
20 cm Aperture Telescope
Two-axis Gimbal
Preamps
Quad Detector Processor
x
y
T1
VPS Controller and Position Detector
T1
yp
xp
CCD Lenses
CCD WFOV Camera
CCD Camera Controller and Pixel to x-y Converter
x
y
T1
BGPS All Digital Controller
x position
y position
Enable
MODE CONTROL ELECTRONICS

FIGURE 6

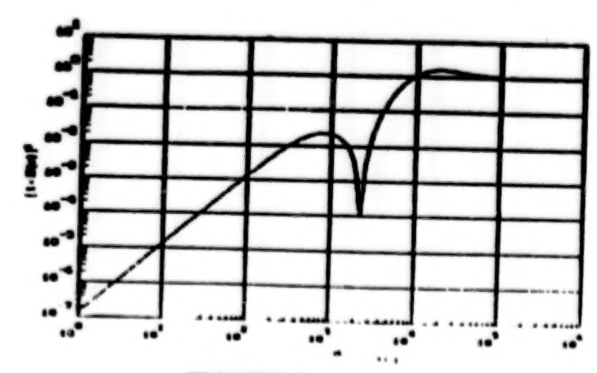


FIGURE 7

ORIGINAL PAGE IS
OF POOR QUALITY

PATS UNTRACKED PLATFORM JITTER

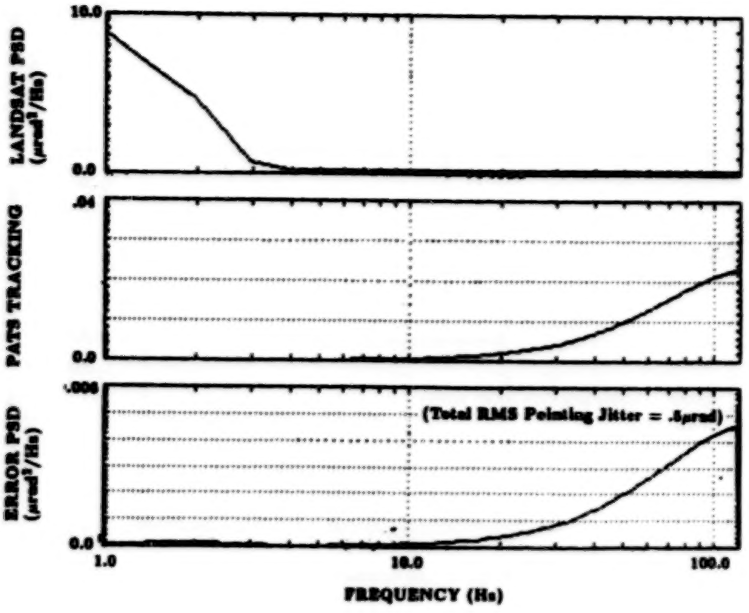


FIGURE 8

SYSTEM DIAGRAM OF 50 Mbps QPPM RECEIVER

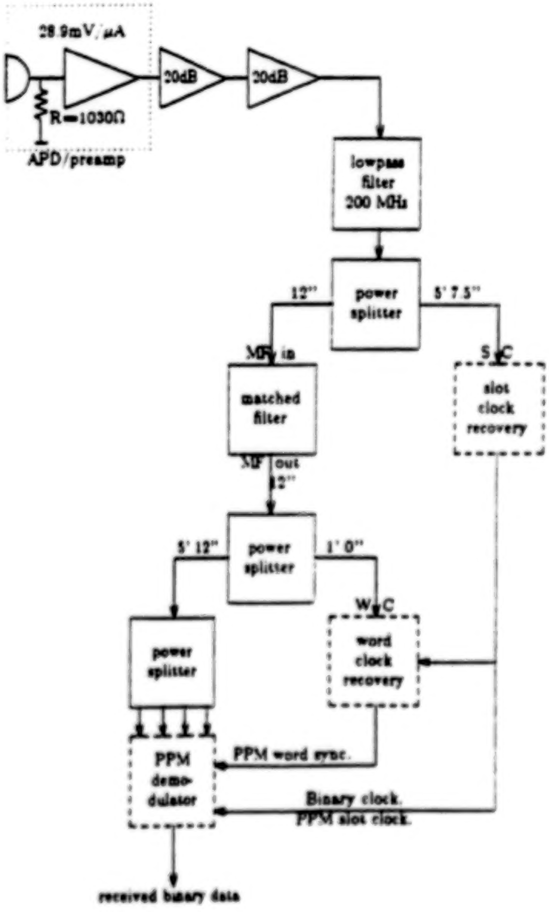

28.9mV/μA
20dB
20dB
R=1030Ω
APD/preamp
lowpass filter 200 MHz
power splitter
12"
5' 7.5"
MF in
S C
matched filter
slot clock recovery
MF out
12"
power splitter
5' 12"
1' 0"
W C
power splitter
word clock recovery
PPM demodulator
PPM word sync.
Binary clock.
PPM slot clock.
received binary data

FIGURE 9

MEASURED AND CALCULATED PERFORMANCE

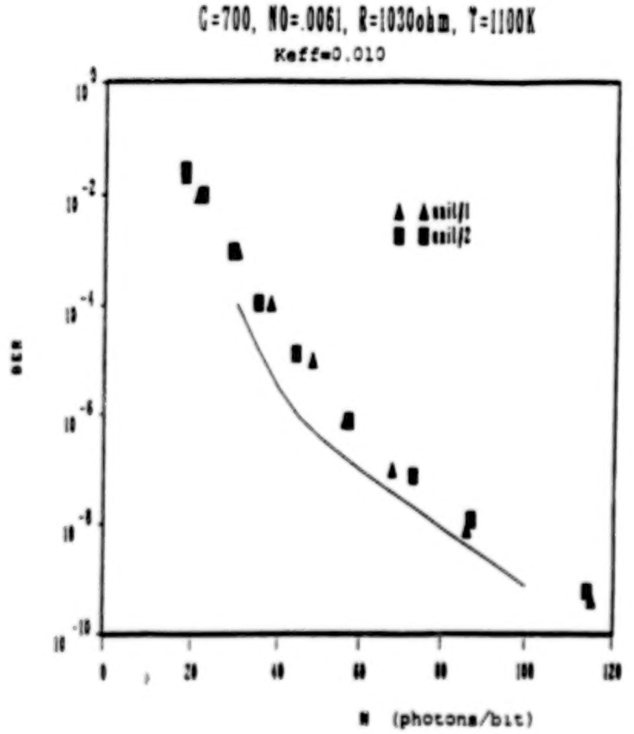


FIGURE 10

ORIGINAL PAGE IS
OF POOR QUALITY

N91

24071

UNCLAS

N91-24071

OPTICAL SHUTTER SWITCHING MATRIX

Charles H. Grove
Chief, CCMS Engineering Section
Mail Code TE-LPS-31
Kennedy Space Center, Florida

VIDEO SWITCHING SYSTEMS

Requirements Overview

Data processing and peripheral support systems, ground telemetry stations, aerospace vehicles, satellites, telecommunications systems, televideo systems, and radar tracking systems are demanding improved interface and switching performance. Today's fiber optic technologies can provide tomorrow's performance systems with the bandwidth, isolation, interference immunity, signal-to-noise-ratio, flexibility, large-scale channelization, and reduction in size, weight, power, and cost.

The National Aeronautics and Space Administration is one of the many government and industry professions which relies heavily on video, data, instrumentation, communications, processors, and tracking systems to perform its mission. The interface switching system technologies needed to meet today's and tomorrow's system performance requirements must be developed and matured if large scale switching systems are to be manufactured that will meet the performance system architectures of tomorrow.

My experience with electronic switching systems is related to those used in the Space Shuttle ground control systems, transmission systems, communications systems, and airborne radar electronic countermeasure systems. My goal is to identify a need that exists throughout the comprehensive information processing and communications disciplines supporting the Space Shuttle and Space Station programs, and introduce one viable approach to satisfy that need. The proposed device is described in the NASA patent entitled, "Optical Shutter Switch Matrix."(1)

Space Shuttle Support Systems

The Kennedy Space Center comprises vast numbers of information and communication systems. The Shuttle Launch Processing System (LPS) is only one of the Center's support systems. The system is made up of three major subsystems. The first is the Control, Checkout and Monitor Subsystem (CCMS) which is the ground support station used to process and launch the Space Shuttle. The second is the Central Data Subsystem (CDS), which provides large-scale data base management for hardware and software configurations, video simulation interface, and real-time data interface used to validate the ground station software. The third is the Record and Playback Subsystem (RPS), which is the ground telemetry station used to record, play back, and analyze ground support equipment and vehicle systems status by evaluating processed and unprocessed information.(2)

The LPS alone supports some 250 engineering work stations, 350 minicomputers, 2 ADP mainframes, 27 computer simulation systems, and over 2,000 peripheral equipment systems comprising magnetic disk, analog and digital tape units, printers, plotters, and graphic recorders. These systems are interfaced through tens of thousands of copper cables, hundreds of patch panels, numerous analog and digital switching systems, and transmission-line conditioning equipment. Together they form the system used to configure, control, test, monitor and record information to and from ground support equipment, and prepare the Space Shuttle for launch. Each console work station contains varying configurations of selectable networks of nonsecure and secure interfaces for multichannel analog and digital voice communications, multichannel operational television, visual displays, and telephones. The local communications centers distribute the secure and nonsecure voice, video, and data to the Centers Satellite Tracking and Data Network and The Air Force Eastern Test Range. The information is then time division multiplexed to NASA, DoD, and Commercial satellite communication networks to other NASA Centers, government agencies, and contractor support organizations.

Shuttle Configuration Requirements

Each Shuttle Vehicle goes through three processing facilities to prepare it for launch. The CCMS firing rooms, CDS, and RPS have to be capable of flexible configurations to support the processing of each Shuttle vehicle in any of the processing facilities. In 1983 the systems were enhanced to support DoD Payload and Shuttle vehicles. The Launch Control Center was equipped with two additional Secure CCMS firing rooms, a Secure CDS, and Secure RPS. The two Orbiter Processing Facilities (OPF), two Vertical Assembly Building (VAB) bays, and two Launch Pads were also modified to support either secure or nonsecure interfaces.

The TEMPEST engineering requirements mandated that extensive modifications be performed to prevent the compromise of information from the existing facilities and interfaces. The existing NASA cabling interfaces, switching systems and communications interfaces could not meet those requirements. Protected cable plant, separate interface cables, large quantities of manual patch panels, and distributed voice, video and data encryption devices were incorporated. The automated switching systems, originally installed to support NASA, and civilian missions, were complemented with optical couplers and manual patch panels to meet the isolation requirements. The total reconfiguration time takes 8-12 hours to configure and validate the 600-700 manual patches required to configure a ground station to support a DoD mission. Large-scale automated switching systems and the technologies needed to develop them were not, and are not, presently available.

In-place Switching Systems

The large-scale switching systems at KSC are based largely on obsolete reed relay technology, TTL Logic, MOS analog switches and amplifiers, digital multiplexers, and logic arrays. Each of these technologies exhibits many of the same inherent electronic design shortcomings. Limited information bandwidth, custom transmission-line conditioning equipment, transmission quality and error-rate considerations, impedance transformations, matching, and long-line equalization, crosstalk, finite input/output isolation, and mutual EMI/EMP/RFI interference are the same concerns still confronting designers. The 200 by 100 and 100 by 100 Remote Control Video Switching systems used in the CCMS are transformer-coupled, three-stage, latching reed relay, cross-bar switches, with operational amplifiers band limited from 2kHz to 2MHz. Each system is enclosed in five standard sized RETMA racks. Rates above or below these frequencies must bypass the switching system and be interfaced through patch panels. Non-zero crossover data types must be converted to a Manchester biphase-L format or connected directly to the equipment through patch panels.

The Remote Control Switching systems in the CDS are designed with digital logic and matching interface circuitry. Switching systems found in the communication systems are designed around MOS analog switches and digital multiplexers with matching input/output interface circuitry. The replacement switching systems being designed today for the next generation CDS are being designed around programmable logic arrays. The improved configurability and signal-to-noise-ratio gained from using this technique is being offset by the 9U board size to accommodate component counts.

Tomorrow's System Requirements

The Center's video, data, instrumentation, communications, processing, and tracking systems at KSC are all near or past their useful design life. System performance, capacity saturation, upgrade limitations, and technology and component obsolescence are the key issues confronting the agency in seeking replacement systems. The goal is to develop design concepts with a 30-year architectural design life. The equipment technologies would be retrofitted as future technology advancements become available.

KSC replacement systems are currently going through their requirements definition, concept design, or initial replacement phases. The users' performance requirements for the replacement systems are driving the next generation systems to the state of the art in optical communication technologies. The success of meeting the requirements will depend on the commercial availability of the technology, maturity, unit cost, installation cost, operations and maintenance costs, upgradability, reliability, and their ability to withstand stress, vibration, and environmental conditions.

OPTICAL SWITCHING SYSTEMS

Requirements Overview

The magnitude, complexity, and problems being experienced in the day-to-day configuration operation of the manual and remote system interfaces supporting multiple commercial and classified missions stimulated the idea for the Optical Shutter Switch Matrix. The daily configurations and system validations of the ground stations are labor-intensive, time-consuming, and prone to error. The major obstacle in automating the configuration process was the requirement to maintain 100 dB isolation between the secure and nonsecure interfaces.

The next-generation information and communication systems for the Space Shuttle and Space Station will use global distributed networks of optical fiber. The communication interfaces being considered include Fiber Distributed Data Interface (FDDI), Broadband Integrated Services Digital Networks (BISDN), and analog and digital Fiber Optic Terminal Equipped transmission lines.

Optical Technology Transition At KSC

The major task for the design agency is to interface the new system architecture into the old system while maintaining compatibility and without impact to the program. Most of the existing facility cable plant was installed in the mid 1960's. No new copper cable plant is being installed. Some fiber optic cables have been installed to free up the existing copper cable plant to meet requirements for the old system. The fiber cable plant that has been activated is being used mainly for TDMA voice communications. The Center's Design organization proposes the complete replacement of the copper cable plant with optical fiber by the mid 1990's. The knowledge and experience with optical transmission systems must be attained quickly to support the major changes which are going to take place at the Kennedy Space Center beginning in 1991. The initial deliveries of systems for the ground stations is scheduled to begin in early 1993. The commercial development and production of optical switching technologies to support the manufacture of large-scale, high performance switching systems needs to be completed by that time.

Optical Switching Concepts

The concept of switching optic-to-optic mediums is not new. Several concepts are available that switch arrays of coupled electro-optic transmitters and receivers, switch optical fibers through mechanical devices, or distribute the optical light waves through directional control mechanisms. The performance of these concepts could exhibit limitations in size, weight, power, switching speeds, reliability, repeatability, expandability, high loss, poor resistance to shock, or high development and manufacturing costs.

Optical Shutter Switch Matrix Patent Number 4,910,396

The need for a multichannel switching technology which met the input-to-output and channel-to-channel isolation requirements, and which would not compromise the integration of both secure and nonsecure information in the same system, prompted the concept for the optical Shutter Switch Matrix. This is the first optical switching concept which combines both mature and newly developed optic technologies into a simple, small, low-power, light-weight, low-cost manufacturing package. The input and output fiber is optically fused to shutter windows mounted in optical glass wafers (Figure 1). The matrix of optical shutter cells are interfaced between a series of orthogonal nontranmissive quartz wafers containing the split and summed optical channels (Figure 3). The individual input channels are power split into a switchable matrix of optical shutter cells, with individual output fiber channels that are power summed. The individual optical shutter cells are controlled by computer addressing (Figure 2). Any input can be switched to any output by electronic or optic activation of the transmissive optical shutter cell (Figure 4). The channel-to-channel isolation is provided in the individual optical wafer, while the input-to-output isolation is dependent on the numb of series shutter matrices incorporated in the package.(1) The device concept is conducive to both single and multimode fiber applications. Switching speeds are limited only by the speed limitations of the types of optical shutter cells employed. Liquid Crystal and Kerr type cells were identified in the initial application because of their maturity, reliability, repeatability, resistance to shock, and reasonable cost.

In June 1989 a Phase I Feasibility Study contract for the Optical Shutter Switch Matrix was awarded to E-TEK Dynamics Inc., in San Jose, California. The study was successfully completed in March 1990. The first 6 by 6 prototype device was manufactured and tested by E-TEK Dynamics to identify any manufacturing limitations, and establish preliminary performance benchmarks. The Phase II Development procurement is planned for early FY-91. An optical switching technology could be ready for production and commercially available to manufacture large-scale switching systems to support the next generation information processing and communications system by 1993.

Phase I Technical Overview

Channel Capacity:	6 by 6 tested 500 by 500 possible
Channel Bandwidth:	≤ 30 GHz
Spectral Bandwidth:	(LC Shutter) 0.6 to 2 micron 2-14 micron possible
Extinction Ratio:	23.1 dB avg. ± 0.85 dB .30 dB possible
Insertion Loss:	≤ 5dB
Switching Speed:	(LC Shutters) < 120 microsec (others tested) 10 microsec to 0.01 nanosec, picosec possible
Features:	Low cost Fabrication simplicity Large spectral bandwidth Large dynamic range Extremely high electro-optic coef. High isolation Excellent expandability Excellent cascadability Excellent repeatability Small size, weight and low power Parallel switching attainable Electronically controlled
Applications:	Large scale optical switches Optical isolators Optical attenuators Optical modulators Optical multiplexers Tunable filters Optical bus interface Network switches for: FDDI, ISDN, FTTH, LAN, WAN, MAN(3)

REFERENCES

1. United States Patent: Grove, C. H.: Optical Shutter Switch Matrix, Patent Number 4,910,396, Dated March 20, 1990.

2. NASA Document: Launch Processing System Description Launch Processing System Division, Kennedy Space Center, November 26, 1984.

3. Final Technical Report: E-TEK Dynamics Inc., optical Shutter Switch Matrix (Phase I) San Jose, California, NASA Contract: NAS10-11550, June 1989 to March 1990.

U.S. PATENT 4,910,396 MARCH 20, 1990
SHEET 1 OF 3

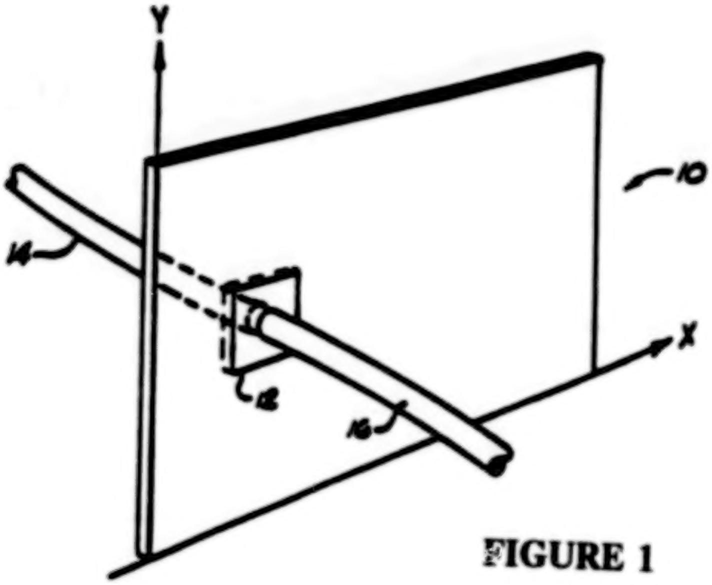

Y
10
14
X
12
16

FIGURE 1

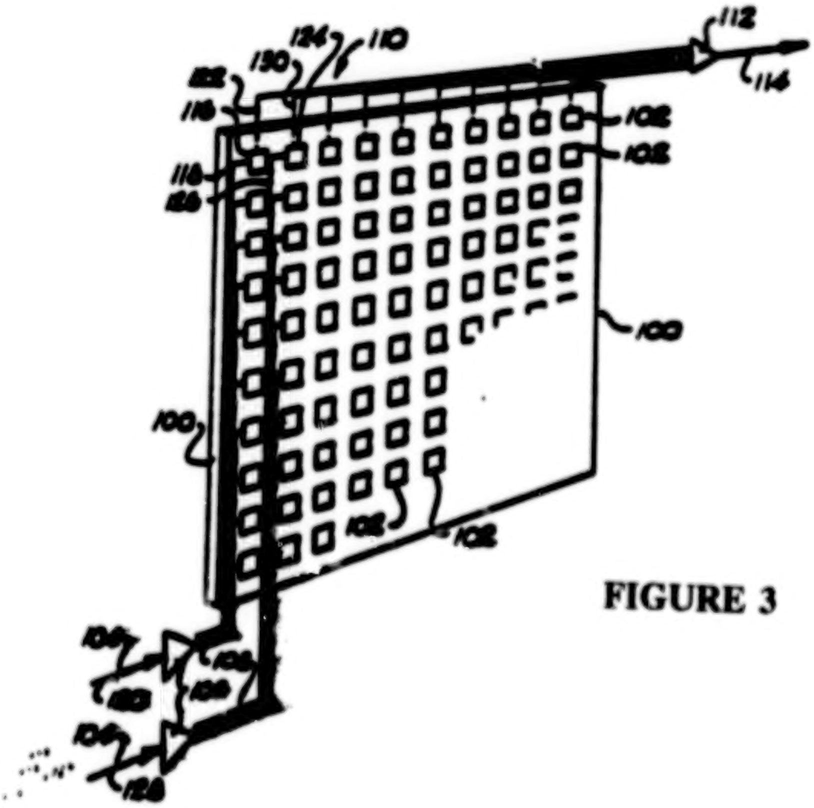

112
124
110
130
122
114
116
102
102
118
120
100
100
102
102
106
126
120
106
128

FIGURE 3

U.S. PATENT 4,910,396 MARCH 20, 1990
SHEET 2 OF 3

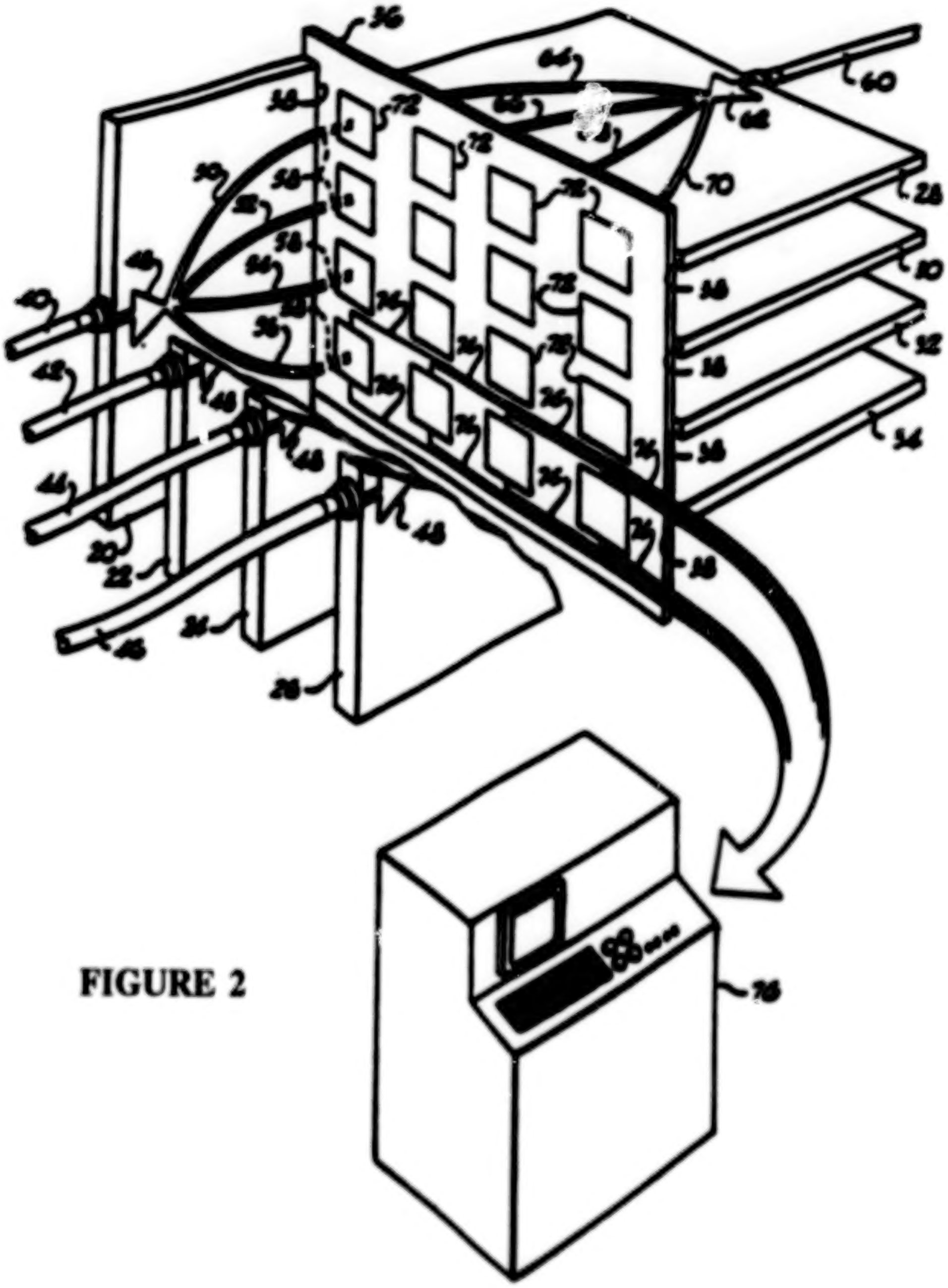

36
64
60
38
72
66
62
68
70
72
50
58
52
28
58
54
30
46
40
56
74
72
38
42
32
72
48
44
48
38
34
74
20
22
48
74
74
38
24
46
26
76

FIGURE 2

U.S. PATENT 4,910,396 MARCH 20, 1990
SHEET 3 OF 3

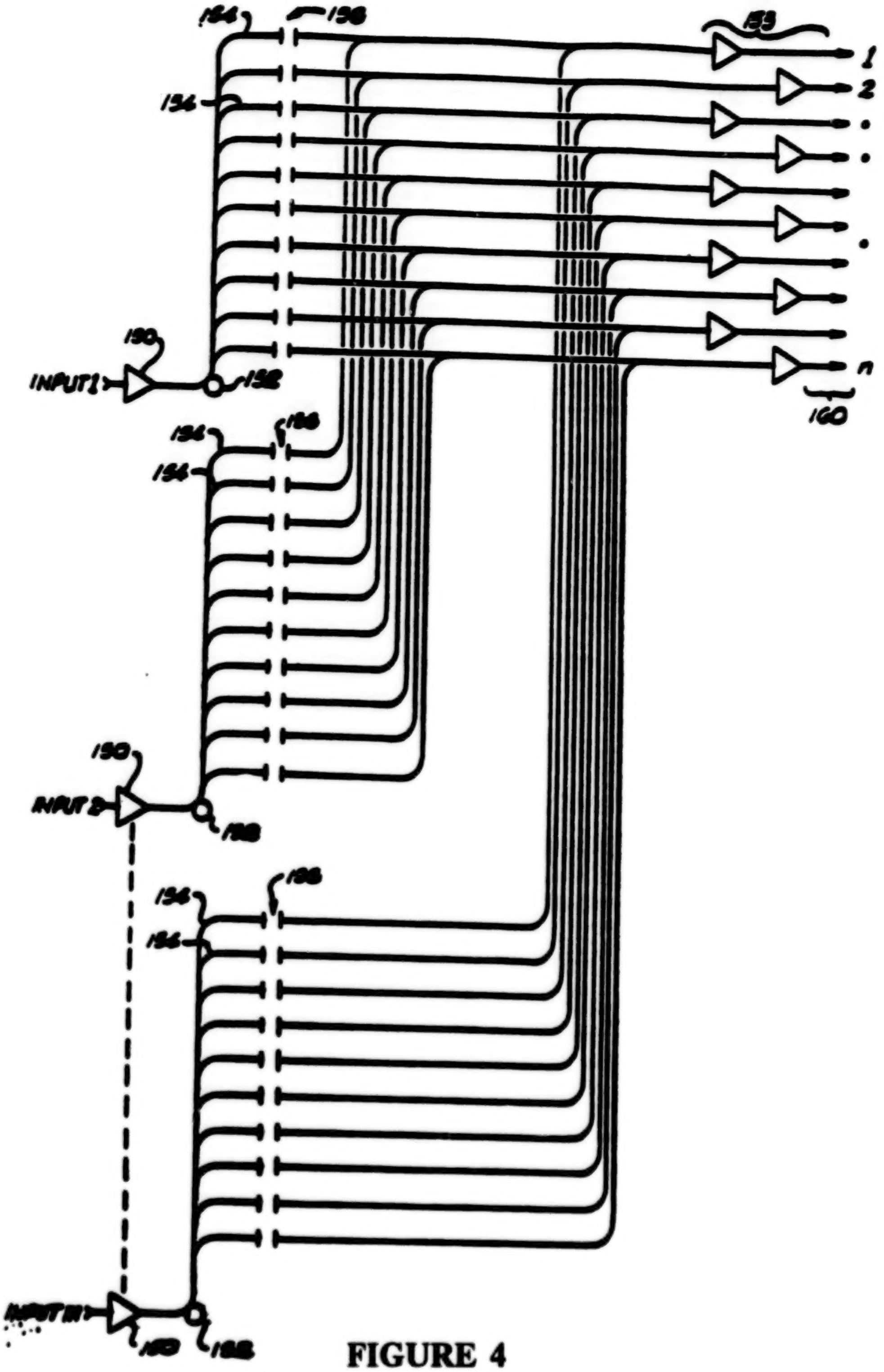

154
158
156
150
INPUT1
152
INPUT2
INPUTm
158
1
2
n
160

FIGURE 4

C4

N91

24072

UNCLAS

N91-24072

Fiber Optic Tactical Local Network (FOTLAN)

L. A. Bergman, R. Hartmayer, W. H. Wu, P. Cassell, G. Edgar,
J. Lambert, R. Mancini, J. Jeng, and C. Pardo
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California 91109

ABSTRACT

A 100 Mbit/s FDDI network interface unit (NIU) is described that supports real-time data, voice and video. Its high-speed interrupt-driven hardware architecture efficiently manages stream and packet data transfers to the FDDI network. Other enhancements include modular single-mode laser-diode fiber optic links to maximize node spacing, optic bypass switches for increased fault tolerance, and a hardware performance monitor to gather real-time network diagnostics.

INTRODUCTION

The FDDI token ring standard provides for a high bandwidth (100 Mbit/s) general purpose interconnection among computers and peripheral equipment using 1.3 μm optical transceivers and 62.5 μm optical fibers as the transmission medium. The network stations are connected in a dual counter-rotating ring configuration with a maximum node-to-node spacing of 2-km [1]. While intended mainly for data traffic, other forms of real-time traffic, such as voice and video, may also be transported over the asynchronous FDDI network using a novel double elastic buffer interface reported earlier in [2] for a lower-speed 80-Mbit/s network. Ordinary data packet traffic was simultaneously transmitted with synchronous T1 voice traffic using a novel voice interface that required only a deterministic network protocol.

Our present effort expands on the previous achievements by migrating the data/voice interface used previously to a FDDI network interface unit (NIU) *and* additionally providing real-time video capability using a technique similar to that used for voice. Furthermore, single mode fiber links, laser diode transmitters and optical bypass switches have been added to increase the distance between adjacent nodes and to introduce network fault tolerance and redundancy. Finally, a network latency performance monitor has been implemented using custom logic.

The NIU system architecture will be described in further detail in Section 2, including the design of the FDDI protocol logic, node processor, voice/video interface, and fiber optic links. The theoretical and experimental performance measurements will be discussed in Section 3, followed by an overview of potential applications in Section 4.

NIU SYSTEM ARCHITECTURE

Concept

The general design provides for computer asynchronous data packets, as well as real-time data packets to be transported by each FDDI node. The major blocks of a typical FDDI node are shown in Fig. 1. The real-time data stream may be accepted from the host computer's bus or more typically from an internal peripheral bus, allowing the operating system not to be degraded by steady real-time traffic loads. FDDI, being an asynchronous network, cannot directly interface to real-time traffic streams without appropriate buffering and resynchronizing logic. In order to interface the real-time voice and video traffic to the LAN elastic buffers, described elsewhere [3], double elastic buffers are added to the input/output sides of each FDDI node with real-time services.

FDDI Chip Set

The NIU design is based on the AMD SUPERNET chip set [4], shown in Fig. 2. The five-chip SUPERNET family meets the ANSI X3T9.5 FDDI standard and acts as an interface between a host computer and the network medium, transferring data and converting it between parallel form at the host and serial form at the media. The chip set consists of the RAM buffer controller (RBC), the data path controller (DPC), the fiber optic ring media access controller (FORMAC), the encoder/decoder (ENDEC), and the ENDEC data separator (EDS). The RBC generates addresses to buffer memory for received and transmitted frames, handles buffer management, and provides interrupts to the Node Processor (NP). The DPC converts data in received frames from byte-wide to 32-bit word formats and vice versa in transmitted frames, performs parity checks, generates frame and node status, and provides interrupts to the NP. The FORMAC performs the media access control layer protocol for the FDDI networking scheme. It determines when a node can get access to the network, implements the logic required for token handling and address recognition, and generates status bits which identify node conditions and frame status. The ENDEC performs the 4B/5B encoding, converts data from parallel to serial format and sends it to the fiber optic transmitter. The EDS extracts the clock from the received bit stream, converts the serial data into a five-bit parallel data format, and performs the 4B/5B decoding.

Node Processor

The node processor (NP) is implemented using a microprocessor based system. Its function is to initialize and oversee the operation of the FDDI chip set, manage the internal NIU bus, perform data conversion/management functions, act as a fault/diagnostics watchdog, and setup external interfaces. Since it functions primarily as a hardware controller, the processor has been designed to respond quickly to packet level interrupt signals. A goal is that NP respond to and process interrupts within a 4μs window. This will allow most real-time interface buffers to be monitored with sufficient accuracy. The 32-bit AMD 29000 RISC CPU [5] was selected from among the several available candidates. It has a high transfer rate, operates at speeds up to 33 MHz, handles DMAC functions, and has strong software and hardware support. The NP operates with a 25-MHz clock input and incorporates four state machines in the design. The first one generates the CPU wait states, the second and third ones provide odd/even word generation and the fourth one provides decoding for the microprocessor's I/O and addressing functions. The NP communicates with the FDDI chip set

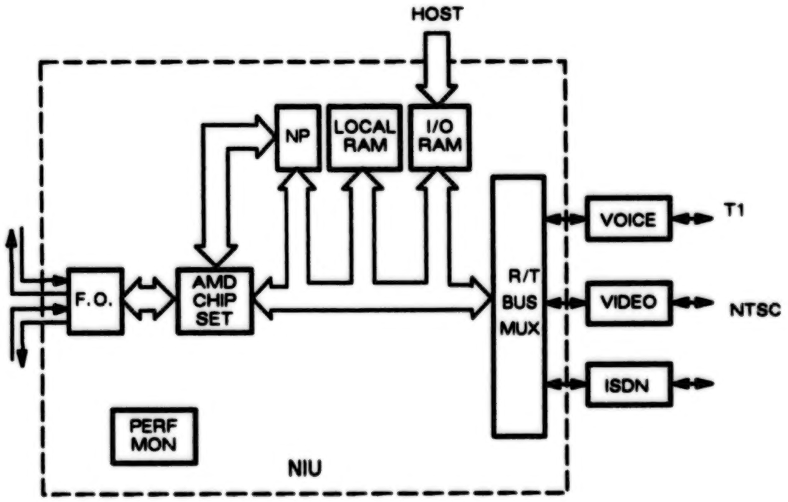


Fig. 1
FDDI node block diagram.

using the 16-bit Node Data and Node Address busses. The SUPERNET chip-set and the buffer memory make their status available to the NP, which has complete control over them.

Voice/Video Interface

The voice and video traffic are both transferred throughout the system in real–time. The voice call setup and switching functions are performed external to the network with PABX equipment [6]. The T1 voice traffic transmitted at 1.544 Mbit/s is stored in the first–in–first–out (FIFO) memory. The 8-bit FIFO busses are then multiplexed and transferred, at 12.5 MB/s, to the NIU data bus via the node path interface circuit. Multiple FIFO channels may be interleaved into one LAN packet. The video signal, fully compatible with the NTSC 525-line, 30-Hz input/output standard, requires a 88 Mbit/s capacity to be transmitted in real–time. Once the pictures are taken with a TV camera, they are sampled at 10 MHz and stored in a 525x525x8 (275,625 bytes/frame) frame store. The frame field position is then added via a header, multiplexed and transferred, at 12.5 MB/s, to the NIU data bus via the node path interface circuit. Since each FDDI packet can only hold approximately 4000 bytes, each video frame is partitioned into 75 packets of 3675 bytes. The frames will be sent via the reliable TCP/IP protocol or via UDP, and corrupted frames will be discarded. In the event that network traffic is high, pictures will be updated less frequently, resulting in "freeze frame" TV pictures. This feature allows the video to act as a free-space bandwidth hog on the FDDI network, giving it considerably more operational flexibility than its voice counterpoint.

Fiber Optic Links

Each FDDI node will be configured in a dual–attach station mode, requiring two fiber optic transceiver pairs and four fiber optic links for a dual counter–rotating network topology. The ring configuration, regular versus loop–back, is controlled by the FDDI chip set. In addition, optical switches allow for the node to be entirely bypassed in the event of optical transceiver failure. Two nodes will utilize laser diode transmitters, extending the distance between them to at least 10 km using single-mode fiber optic cable.

Transceivers. The fiber optic transceivers for normal FDDI operation, are hybrid modules [7] operating at 125-Mbit/s NRZ. The transmitter utilizes a high–speed LED having an average rise/fall time of 2 ns and output power of –15 to –20 dBm at λ=1305–1380 nm. The receiver uses a PIN

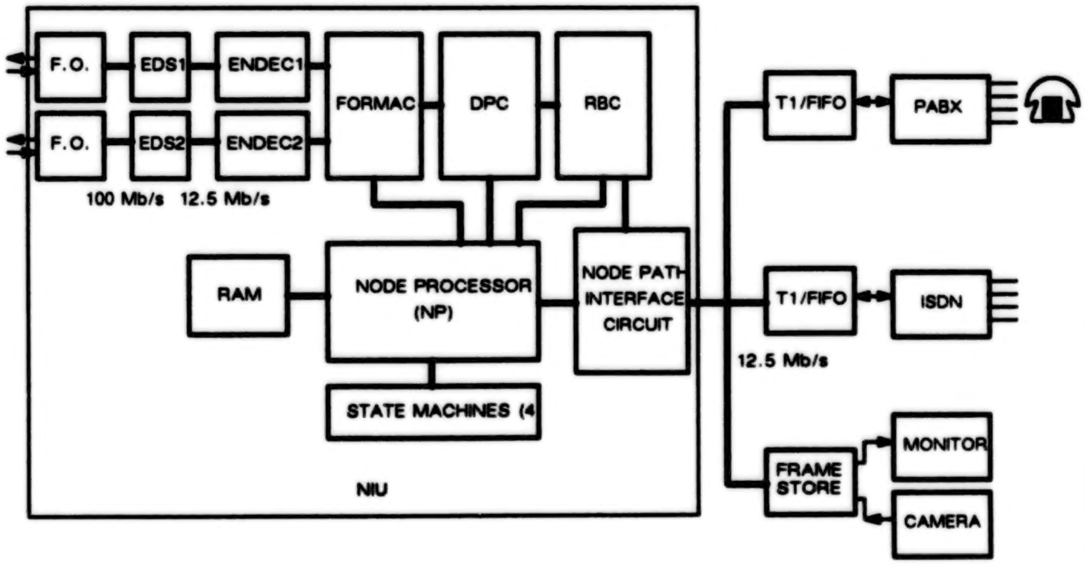


Fig. 2
Detailed FDDI node/system block diagram.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE IS
OF POOR QUALITY

Fig. 3
Fiber optic cable assembly.

photodiode and has a sensitivity of −34 dBm [8]. Both the transmitter and the receiver are terminated with an ST type connector. For extended distances greater than 3 Km, a single-mode laser diode transmitter, having an output power of 0 to −3 dBm will be used. The receivers, having sufficient sensitivity, remain unchanged.

Dual Fiber Cable. The ruggedized two-fiber tactical fiber optic cable contains two tightly buffered 50/125-μm multi–mode optical fibers [9] and are terminated with hermaphroditic biconic connectors [10], which facilitate field deployment and retrieval without physical or optical degradation. The fibers are radiation hard and fully militarized to withstand the tactical field environment. The fibers are cabled in a ruggedized all-dielectric structure with kevlar yarns and surrounded by reinforcing elements and an outer jacket. Their attenuation is 1.0 dB/km at 1300 nm. The outside diameter is 6 mm and its weight is 30 kg/km. The link optical power budget is 10 dB, yielding a maximum transmission distance of 3 km. A typical fiber optic cable assembly is shown in Fig. 3.

Single Mode Fiber Cable. The long distance transmission links will utilize ruggedized single mode tactical fiber cables [11]. These cables, similar in packaging to the multi–mode version, contain a radiation hardened, single mode optical fiber. The fiber has a core diameter of 8.3 μm and has an attenuation of 0.35 − 0.50 dB/km at λ=1310 nm. At this wavelength, which is also the zero dispersion wavelength, the maximum fiber dispersion is 3.2 ps/nm–km.

Optical Bypass Switches. The optical bypass switches enable the network to be reconfigured in two different ways. A single–node will be bypassed in the event of an optical transceiver failure. The data will enter the node and exit immediately, without reaching the FDDI card. In the second bypass mode, an entire portion of the ring can be disconnected, thereby bypassing several nodes. This scheme would be useful if a larger portion of the ring is not functional or is even destroyed. Ideally, a low attenuation 3x3 optical crossbar switch would be used in order to implement this scheme. Since no commercial product is available at this time, two 2x2 switches connected in series are being used. Since the increased loss exceeds the available power budget, the optical bypass will only be used with the laser diode transmitter links.

Packaging

The FDDI boards will be implemented on a PCB, which will conform to the VME standard. For the prototype version, the design will be broken down into three 6U high boards operating in a standard VIM/VME cardcage. The final version will include all the different FDDI functions on one 9U high card, which can be used in a SUN workstation setup [12]. Test points for signal monitoring will be included, as well as jumpers isolating the separate board functions, if needed. LED indicators will also be included to monitor the link operation.

PERFORMANCE

Theory

The predicted network performance is illustrated in Fig. 4 [13]. Time delay versus throughput for several token protocols are shown. Observe that FDDI offers a capacity of 768 voice channels at time delays of ~1ms or more. In practice, fewer voice channels would be used to give reasonable access delays for data-only packets. For example, in a network file system (NFS) service, an access delay of <250 μs would be desired thereby restricting the throughput to about 30 Mbit/s or 240 voice channels. The capacity of low-speed LANs is too low to support T1 traffic. However, single voice channels might be supported using the same technique with 802.4 or 802.5 LAN protocols.

Simulations

Real and Non-Real-Time Performance Estimation. A simulation of the data flow within the NIU can provide valuable feedback to designers on how to optimize the overall hardware performance. For example, to maximize end-to-end throughput one might intuitively allocate an equal bandwidth among each of the functional blocks (e.g., bus, memory). However, in a system where real-time and non-real-time traffic are freely mixed together, the message sizes, frequencies, and statistical distributions will vary dynamically, thereby making such an estimation considerably more difficult. In this instance, packets may be forced to be randomly stored (or buffered) at numerous locations creating bottleneck points. For example, a low bandwidth introduces high queue populations for even moderate network/host loads and consequently high latencies. Conversely, a very high-bandwidth maintains a low-Q population but at very low internal bus utilization, and at great cost. Thus, an optimum exists.

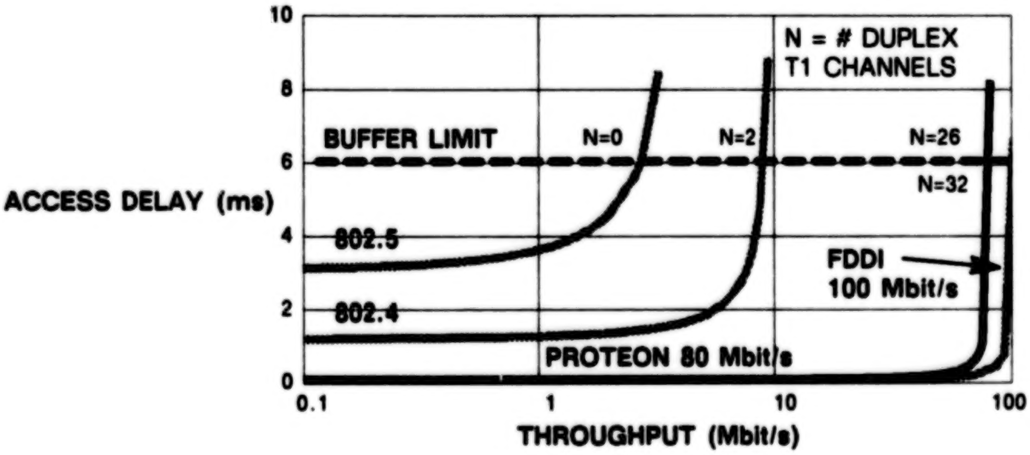


Fig. 4
Access time delay is plotted as a function of throughput for four different types of token ring LANs. The 6-ms limit represents the current voice buffer size. Observe that longer delays permit more rescheduling flexibility and thus more effective bandwidth utilization.

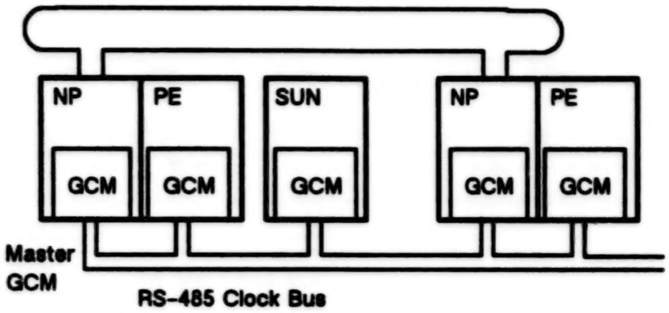


Fig. 5
Block diagram of the Global Clock Module (GCM).

Discrete Event Simulator. An object-oriented graphical discrete-event modeling environment such as *SES/Workbench* can provide an effective means of estimating the minimum buffer depths and bus bandwidths to control queue buildup and excess latency. The objective was to study the effect of buffer memory bandwidth on overall performance throughput and latency, and suggest reasonable performance goals to the designers. The RAM buffer design suggested by AMD has a bandwidth of 200 Mbits/s. In the case of input distributions of 3T1, 2 ISDN, 2T2 and Poisson processes for computer data, the simulation results show that a buffer memory bandwidth of under 200 Mbits/s can cause a severe bottleneck (as expected), but that a buffer memory bandwidth of around 300 Mbits/s is optimal (and less obvious). If the portion of the network bandwidth allocated to video is increased from 12 Mbit/s to 40 Mbit/s, a very high Q population for communication specific bus will result even with a bandwidth of 350 Mbit/s for buffer memory. Thus, even larger bus/buffer memory bandwidths may be desirable in these cases.

Measurements

Performance Monitor. The purpose of the performance monitor is to provide a means to measure the end to end network latency and throughput at the TCP and IP layers, and the queue timing at the NP. In measuring these parameters, we need to minimize the influence of the protocol as well as the software processing. In order to synchronize timing measurements, a global reference time needs to be established throughout the LAN. Each NP module will include a global clock module (GCM), which will read a reference time with 100-ns resolution from an external common reference source. One master GCM will orchestrate all timing measurements throughout the network and synchronize the reference time. The real-time measurements will be stored in a FIFO, to be collected by the CPU at its convenience. A block diagram of the GCM distribution is shown in Fig. 5.

Workstation Throughput. The throughput is measured with the User Datagram Protocol (UDP) and Transmission Control Protocol/Internet Protocol (TCP/IP) provided the workstations [12]. UDP is considerably more efficient, but does not perform flow control, acknowledgments, and error checking. The test software generates long blocks of random data in RAM and directly sends this to the network, bypassing the disk subsystem. At the receiving workstation, the same data is discarded on a byte-by-byte basis.

For a TCP/IP (non-Van Jacobson) protocol and unloaded LAN, the throughput between two workstations was measured to be 2.2 Mbit/s for a 80 Mbit/s LAN and 1.8 Mbit/s for an Ethernet network (Sun 3/160 workstations). The 20% improvement is limited by the TCP/IP software overhead. For the simpler UDP protocol, the token ring provides a two to three-fold improvement over Ethernet. Still, the throughput is much less than expected from theory—a limitation imposed by the operating system overhead and hardware platform. Throughputs exceeding 30 Mbit/s is expected between current RISC-based workstations.

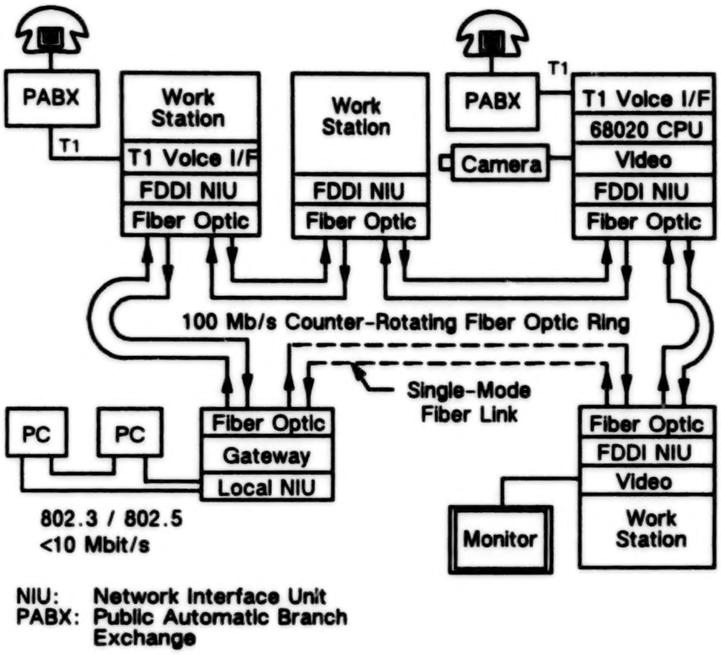

PABX
T1
Work Station
T1 Voice I/F
FDDI NIU
Fiber Optic
Work Station
FDDI NIU
Fiber Optic
PABX
T1
T1 Voice I/F
68020 CPU
Camera
Video
FDDI NIU
Fiber Optic
100 Mb/s Counter-Rotating Fiber Optic Ring
Single-Mode Fiber Link
PC
PC
Fiber Optic
Gateway
Local NIU
802.3 / 802.5
<10 Mbit/s
Fiber Optic
FDDI NIU
Video
Monitor
Work Station
NIU: Network Interface Unit
PABX: Public Automatic Branch Exchange

Fig. 6
Fiber optic local network testbed.

Voice. Bit error rate (BER) tests are performed at the T1 interface of two units attached to the ring. Error injection tests indicate that BERs on the order of 10^{-4} were not discernible.

APPLICATIONS

Army. The testbed, shown in Fig. 6, will be used to demonstrate the feasibility of serving a variety of Command, Control, and Communication (C^3) needs within the tactical Army using a FDDI based LAN for high-speed integrated voice/video/data communications. Modular network concepts, rapid deployment capability with fiber optics, and compatibility among several types of Army stations and LAN programs have already been investigated [14].

NASA: The techniques developed on FOTLAN for streaming voice/video traffic over a packet network may also be used many future NASA applications in both the spaceborne and terrestrial arenas. Spaceborne examples include spacecraft data buses for combining packet data and high-rate instruments such as HIRIS and SAR onto one fiber optic media. Within a decade, other stream-oriented instruments may also be added to space vehicles such as optic storage systems, optical processors, and systolic array processors. One of the important advantages of such a network is that it would be rapidly reconfigurable, making it relatively easy to relocate and replace subsystem modules in spacecraft assembly or for repair/upgrading by the shuttle. For ground based applications, the rapid deployment capability of FOTLAN may be important in simplifying cable layout for data/voice/video in some mobile tracking stations.

CONCLUSIONS

With the emergence of FDDI compatible components, it is only a matter of time before high-speed networks will be in common use. While the conventional FDDI standard only requires multi-mode data transmission over 2 km spans, the single-mode fiber and real-time interface augmentations offered here bring the standard to a more advanced plateau with added functionality, more diverse applications, and broader geographical coverage. Finally, the multi-level communication fabric proposed here provides, for the first time, a basis for a single integrated service workstation with seamlessly merged voice, video, and data graphics capability. Such end applications as voice/video E-mail, multi-station video conferencing, and distributed tactical command posts are but a few of the possibilities. We expect stream-oriented ISDN to eventually provide similar capability, but its hierarchical fixed address topology may limit its usefulness in environments where rapid and frequent network reconfiguration is a requisite.

ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions of team members G. Bolotin and L. J. Paul.

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by the U. S. Army CECOM through an agreement with the National Aeronautics and Space Administration. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government, the U. S. Army CECOM, or the Jet Propulsion Laboratory, California Institute of Technology.

REFERENCES

[1] FDDI 2 (X3T9.5) Working Paper, Feb.1, 1987.

[2] Bergman, L. A.; Hartmayer, R.; Marelid, S.; Wu, W. H.; Edgar, G.; Cassell, P.; Mancini, R.; Kiernicki, J.; Paul, L. J.; Jeng, J.; Pardo, C.; Halloran, F.; and Martinez. J.: A Fiber Optic Tactical Voice/Data Network Based on FDDI. Armed Forces Communications and Electronics Association (AFCEA) DoD Fiber Optic Conference '88, McLean, VA, pp. 123–125, March 22–25, 1988.

[3] Bergman, L. A.; Hartmayer, R.; Marelid, S.; Wu, W. H.; Edgar, G.; Cassell, P.; Mancini, R.; Kiernicki, J.; Paul, L. J.; Jeng, J.; Pardo, C.; Halloran, F.; and Martinez. J.: A Fiber Optic Tactical Voice/Data Network Based on FDDI. INFOCOM'88, New Orleans, LA, pp. 45–54, March 29–31, 1988.

[4] Advanced Micro Devices – The SUPERNET Family for FDDI. [EDS: Am7984A, ENDEC: Am7985A, FORMAC: Am79C83, DPC: Am79C82, RBC: Am79C81]

[5] Advanced Micro Devices – Am29000 32-bit streamlined instruction processor.

[6] Wang WBX public automatic branch exchange.

[7] AT&T ODL-125TR (105711923) Light Wave Data Links.

[8] Specified at 125 Mb/s data rate, using 62.5 μm fibers.

[9] AT&T 10524331 tactical dual-fiber cable.

[10] AT&T 105243596 dual-fiber bulkhead connectors.

[11] AT&T radiation hardened fiber.

[12] Sun Microsystems, Inc., 3/160M.

[13] Bergman, L. A.; Halloran, F.; and Martinez, J.; Implementation of a Tactical Voice/Data Network Over FDDI. MILCOM'88, San Diego, CA, October 23-26, 1988

[14] Frankel, M. S.; et al: An Overview of the Army/ DARPA Distributed Communications and Processing Experiment. IEEE J. Select. Areas Communications, Vol. SAC-4, No. 2, pp. 207–215, Mar 1986.

N91

24073

UNCLAS

N91-24073

HIGH PRECISION APPLICATIONS OF THE GLOBAL POSITIONING SYSTEM

Stephen M. Lichten

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

ABSTRACT

The Global Positioning System (GPS) is a constellation of U.S. defense navigation satellites which can be used for military and civilian positioning applications. The GPS will ultimately include at least 21 satellites at about 20,000 km altitude equally spaced in six orbit planes by the early 1990s. A wide variety of GPS scientific applications have been identified and precise positioning capabilities with GPS have already been demonstrated with data available from the present partial satellite constellation. Expected applications include: measurements of Earth crustal motion, particularly in seismically active regions; measurements of the Earth's rotation rate and pole orientation; high-precision Earth orbiter tracking; surveying; measurements of media propagation delays for calibration of deep space radiometric data in support of NASA planetary missions; determination of precise ground station coordinates; and precise time transfer worldwide.

INTRODUCTION

The Global Positioning System (GPS) is designed so that typically six to ten navigation satellites can be tracked above 5 deg elevation from any ground site. The GPS satellites transmit carrier signals at 1.227 and 1.575 GHz (L-band) which are modulated by a pseudorandom noise code, the P-code, at 10.23 MHz. Two frequencies are provided so that ionospheric signal delays can be calibrated. A second code, the C/A (coarse acquisition) code, is somewhat noisier than the P-code due to its lower frequency at 1.023 Mhz and the lack of dual-band ionospheric correction (see Fig. 1)

The GPS codes include a navigation message with GPS clock and orbit information which can be utilized for real-time point positioning by users equipped with GPS receivers. GPS *pseudorange* to four satellites determines three position coordinates plus the user clock offset from GPS time (Fig. 2). The term pseudorange is used since the range calculation is based on the difference between the transmit and receive times and will include any offset between the transmitter and receiver clocks. With the P-code, user positions can be determined in near-real time to about 10 m.

The Department of Defense plans to turn on *selective availability* (SA) for the operational GPS satellites. SA includes a clock dither and changes to the broadcast ephemeris. Authorized users will be equipped with keys to correct for these effects. Other users will see transmitter clock variations of the order of 30-50 m and the broadcast ephemeris will be accurate to 50-100 m as well. Few scientific and non-military GPS users will be equipped with the keys for selective availability. Simultaneous GPS tracking from multiple receivers can differentially eliminate GPS and receiver clock offsets. This method of differential GPS tracking is appropriate for non-real time applications for which simultaneous data from distant sites must be brought together and combined. Because the broadcast ephemeris provides only coarse orbit information, for higher precision the user must estimate and improve the GPS orbits *and* use differential GPS techniques (Fig. 2).

The primary data type for highest precision in non-real time GPS applications is the carrier phase, which can be tracked with sub-cm precision in modern GPS receivers. The carrier phase, continuously tracked over hours, provides a precise time history of *range change* which can be used with a dynamical model to obtain precise orbit solutions (Fig. 2). The GPS carrier phase is ambiguous by some integer multiple of carrier wavelengths, but changes in the phase from point to point measure range change. Pseudorange data, even when it is some orders of magnitude noisier than the phase, is still useful since it can be used to constrain the carrier phase integer ambiguities, which in general must be estimated from the data.

The Department of Defense may also encrypt the P-code data (anti-spoofing, or AS), making it unavailable without special devices. There are several methods discussed below which enable recovery of a noisier pseudorange data type using codeless techniques when AS is on. Although AS is presently generally off except for intermittent testing, with the full operational GPS constellation it may be turned on routinely.

With AS *off*, a number of commercial receivers can provide dual-frequency calibrated pseudorange data with intrinsic receiver system noise of tens of cm over five minute averaging intervals. Most of the actual observed scatter, however, is much higher, about 50-200 cm, and is due to multipath. The Rogue digital GPS receiver [1] developed at the Jet Propulsion Laboratory (JPL), incorporates both a very low noise pseudorange system noise (< 5 cm) and an antenna with choke rings designed to minimize multipathing. Fig. 3 shows samples of data demonstrating the low noise capability of this instrument. Studies have shown that low noise pseudorange is beneficial in high precision positioning applications.

With AS *on*, GPS users without the classified decryption keys can still make precise measurements with carrier phase provided they use codeless receivers. There are several different types of codeless receivers. The squaring-type of codeless receiver uses the fact that the P-code [P(t) in Fig. 1] has +/-1 values in a pseudorandom pattern whose square is always +1. An alternative approach is used in the Rogue receiver [2], which cross correlates the L1 and L2 signals and relies on the C/A signal together with the P1-P2 delay (from cross correlation) to produce carrier phase as well as pseudorange with an ionospheric correction. Codeless Rogue carrier phase is much noisier than code tracking phase, as is the pseudorange. For example, P-code pseudorange with the Rogue has receiver thermal noise at the cm-level, while the codeless pseudorange has noise at the 20-30 cm level [2]. Additional multipath errors tend to reduce the contrast between code and codeless pseudorange. The Rogue receiver defaults to P-code tracking but reverts to codeless mode as soon as it detects that AS is on.

The cross-correlation type of receiver has several advantages over the squaring receivers. First, the effective wavelength is not changed by the cross correlation, while squaring codeless receivers result in an effective wavelength one-half the original ~ 20 cm L-band wavelength. The shorter wavelength makes phase connection, data editing, and phase ambiguity resolution more difficult since all these procedures rely on determination of signals and parameters to a fraction of a wavelength. Secondly, the cross-correlation receiver such as the Rogue can produce an ionospherically calibrated pseudorange from the C/A signal and L1-L2. Although this pseudorange may be somewhat noisier than P-code pseudorange, it can be considerably more accurate than pseudorange from codeless receivers which rely only the single frequency C/A signal.

Future government policies will determine in part how effective different GPS receivers are for unclassified applications. With AS off, GPS receivers which fully exploit the P-code have the advantage. However, since in the future AS may be turned on, a backup codeless capability is imperative for maximum accuracy.

GPS receivers and antennas are relatively small and portable. They can be transported to and set up in remote locations by one person and can be run from batteries or small generators. A number of different code and codeless models are presently available from various manufacturers. The market for GPS technology has grown enormously over the past few years. Although early models were relatively expensive, modern high-precision GPS ground terminals can be purchased for well under $100,000 and lower-accuracy receivers cost under $5,000. Inexpensive GPS receivers are now available for installation in automobiles, with accuracy claimed to about 100 m [3]. Space-qualified GPS receivers for Earth orbiters require on the order of $1 million or more, but production costs are expected to drop after component parts are reduced.

HIGH-PRECISION GPS TRACKING

Least Squares Estimation

Many high-precision GPS applications require estimation of GPS orbits and other parameters. With SA turned on, the GPS broadcast ephemerides are accurate to approximately 50 m. On the other hand, geodetic experiments designed to measure Earth crustal motions must detect ground motion at the cm or even sub-cm level from one year to the next; this in turn, requires sub-meter GPS orbit accuracy. For global geophysical studies or precise low-Earth orbit determination, GPS orbit accuracy of 10-20 cm is the goal.

Many high-precision GPS scientific applications therefore involve studying ways to determine GPS orbits to very high accuracy. This is a least squares estimation problem to which square-root Kalman filtering and smoothing techniques are well suited [4]. Differential GPS tracking is used to remove errors due to transmitter and receiver clocks or selective availability. This requires combining GPS data from different receivers to estimate parameters which include: GPS orbits and solar pressure coefficients, ground site coordinates, biases and clocks, and tropospheric signal delays. Considerable research at JPL has led to a GPS data processing strategy which incorporates a number of refinements designed to improve accuracy. These refinements include improved physical modeling of the Earth and of the satellite motion, and stochastic representation for atmospheric signal delays. The JPL software also includes a rapid, automated data editor [5], and incorporates specialized techniques to resolve carrier cycle ambiguities [6]. The launch of the Topex/POSEIDON oceanography satellite in 1992 into low-Earth orbit (1336 km altitude) will result in the first high-quality GPS flight data from the GPS experimental receiver on Topex. The techniques for orbit determination of a satellite using GPS data collected with a flight instrument are similar to those which have been used to determine ground station coordinates with GPS. The main difference is that the solution for an orbiting receiver is influenced by force mismodeling, which could include errors in the gravity field, solar radiation pressure model, drag coefficient, etc. The *reduced-dynamic tracking* strategy [7] is a filtering technique developed at JPL to minimize or eliminate such force mismodeling errors. Stochastic acceleration parameters are estimated to absorb dynamic mismodeling. The constraints on the stochastic parameters are chosen to correspond to the level of confidence in the dynamic models. When reliable dynamic models are available (e.g. for GPS satellites themselves in high-altitude, circular orbits), very tight constraints are placed on the stochastic acceleration parameters, while for satellites in low-Earth orbit where dynamic modeling is poorer due to geopotential uncertainty, the stochastic acceleration parameters are loosely constrained so that more weight is placed on GPS tracking data to fit out unmodeled forces. With this technique, GPS-based orbit determination for Topex is expected to be accurate to the 10-cm level or better [7].

Current research in least-squares estimation techniques for GPS applications includes approaches to reducing computation. With the full GPS constellation and dozens of ground tracking sites, the computational burden of solving for hundreds of parameters will be significant. Determination of geopotential parameters may involve thousands of parameters and even supercomputers may be inadequate with a standard least-squares approach. Recent studies at JPL have identified new approaches to matrix partitioning and decentralized processing which show promise for reducing computation time for these demanding GPS tracking applications by more than an order of magnitude [8, 9].

Ground Positioning Applications

Ground positioning GPS applications can be classified into three categories: (1) global geodetic and geophysical studies, including measurements of intercontinental distances, Earth rotation rate, and position of the Earth's pole to an accuracy of a few cm; (2) regional geodetic measurements of crustal motion and deformation oriented towards sub-cm level ground measurements over distances ranging from ~ tens of km to 1000 km; (3) surveying applications, which are usually local measurements over relatively short distances. The first category is among the most demanding in terms of GPS orbit accuracy due to the long distances (thousands of km) involved. The second category also requires highly precise orbital modeling. The third category generally requires less accuracy for the GPS orbit solutions than the other geodetic applications.

Global Measurements

The determination of the Earth's variable rotation rate is very important for astronomical and deep space tracking applications. Also important is the determination of the rotation axis of the Earth relative to the Earth's crust, which moves back and forth relative to the pole. Presently, the determination of Earth orientation is made with very long baseline interferometry (VLBI) measurements using large, fixed radio telescopes; with satellite laser ranging (SLR); or with lunar laser ranging (LLR). Although the VLBI, SLR, and LLR Earth orientation measurements are accurate to the level of a few cm, the results are available on a routine basis with resolution of only several days and a delay of a few weeks. A better understanding of Earth rotation requires higher time resolution to better than 1 day, and to properly calibrate NASA deep space tracking data, the results should be available within 24 hrs. Analysis [10] indicates that GPS tracking from a worldwide network with the full satellite constellation can provide daily or better time resolution for cm-level changes in Earth orientation, thus complementing nicely the precise but less frequent absolute measurements available from VLBI. The incorporation of GPS data could reduce the demand for Earth orientation calibrations using VLBI with NASA's deep space network (DSN), while at the same time improving the accuracy of these calibrations. Initial pole position measurements with GPS [11] are encouraging, with agreement with VLBI demonstrated at the level of 5-10 cm. Further improvement to the cm-level is anticipated in the next few years as additional GPS satellites are launched. Other global measurements expected from GPS with few-cm accuracy include determination of the geocenter (Earth center of mass) and long baselines. Measurements of long baselines, for example over intercontinental distances of thousands of km, will be useful in studies of large scale plate motions and continental drift. Presently VLBI is the most accurate technique available for making these measurements; however the relatively small and portable GPS terminals will enable many long baseline measurements to be made which had previously been unavailable due to the limited number VLBI and SLR observatories. Initial GPS solutions for intercontinental baselines determined from the limited configuration of a worldwide 1988 GPS experiment show good agreement — about 5 cm over 5,000-6,000 km, or 1 part in 10^8 — with independent VLBI solutions for the same baselines [11, 12]. This experiment will be repeated in 1991, by which time the number of GPS satellites and ground tracking sites will have more than doubled as compared to 1988 and solutions should improve to several parts in 10^9.

Regional Measurements

Regional geodetic measurements are usually made over distances of hundreds of km. Hundreds of GPS receivers have been deployed worldwide in numerous regional geodetic experiments in the past few years, and GPS is well established as a new and important high-precision technique to measure crustal motion. Present-day GPS errors for relative horizontal ground measurements include a system noise component of a few mm which does not depend on the baseline length being measured, and a baseline length-dependent error component of about 1 cm per 1000 km of length (1 part in 10^8) [13]. Although most plate motion is horizontal, vertical displacements can also be determined with GPS, although with somewhat less accuracy (~ 3 cm level). There are many high-quality GPS geodetic results which have been published in the literature. In this paper, I will present three recent samples of results.

1. *Measurement of volcanic fracture opening in Japan with GPS.* A recent paper [14] presents results of monitoring a seismically active region in Japan with GPS. In July 1989, there was a swarm of earthquakes, the largest of which was magnitude 5.5, near the Izu Peninsula. Just after these earthquakes, the GPS solution indicated ground motion of nearly 15 cm had occurred. Shortly thereafter, a volcano nearby on the sea floor erupted. The crustal movements detected using GPS were interpreted as a fracture opening around the swarm region. The fracture of the crust was associated with upward motion of magma. The Japanese GPS experiment demonstrated that GPS measurements can reveal the time sequence of seismic and volcanic events.

2. *Measurement of plate motion along the San Andreas fault in California.* The San Andreas fault in California marks the boundary of the Pacific and North America plates. The Pacific plate is moving northwest relative to the North America plate at approximately 5 cm/yr. In order to understand the risk of earthquakes and their relation to plate motion and deformation, numerous GPS sites have been occupied in California for several years. Fig. 4 shows one example of the measurements recently made using GPS [15]. The two ends of the baseline are on opposite sides of the San Andreas fault. The plate motion is clearly detected and agrees with independent measurement techniques to the level of a few mm/yr. By making measurements such as these annually throughout tectonically active regions such as California, a better understanding of the connection

between earthquakes and the rates of plate motion and associated deformation can be reached. The only feasible technique for occupying dozens of ground sites along the complicated fault networks in California is use of portable GPS receivers.

3. *MM-level precision for a geodetic measurement.* A recent paper [16] presented measurements of ground baselines in California of length 245 and 729 km. The measurements were repeated daily over a 1-week period in 1988. In the horizontal components, the daily solutions agree with a rms scatter of 1-4 mm. The GPS orbits determined simultaneously are believed to be accurate to about 50-100 cm [17]. These results indicate the high precision which can be achieved with GPS techniques.

Surveying Applications

Surveying applications abound for GPS. One advantage of GPS is that areas can be surveyed even when there is no clear line of sight between pairs of sites. In an interesting example of a GPS surveying application [18], a GPS satellite survey was utilized to support construction of the Stanford Linear Accelerator Center (SLAC). The accelerator plans called for a loop ~ 1 km in diameter. Approximately 1,000 magnets had to be placed in the arc tunnels, requiring a network of reference marks. The GPS survey, using codeless GPS receivers, revealed systematic errors in the terrestrial solution, and was reported to be accurate to 1-2 mm horizontally and 2-3 mm vertically [18]. Other users of GPS for surveying include the Pennsylvania Department of Transportation, Department of Lands and Survey of New Zealand, the Canadian Geodetic Survey, and the U.S. National Geodetic Survey (see ref. [18], proceedings to a GPS conference in which numerous other surveying applications were discussed). Surveying applications have less demanding GPS hardware and orbit accuracy requirements since relatively short distances are typically involved. Over shorter distances there is greater cancellation of errors from the atmosphere and from GPS orbits.

Orbiting GPS Users

The illustration in Fig. 2 shows schematically differential GPS tracking for an Earth orbiter equipped with a GPS flight receiver. Differential positioning requires a ground network of GPS receivers. This technique will be used for the Topex GPS demonstration [7]. Topex/POSEIDON will be launched in 1992 and will carry an altimeter precise to about 2 cm to map the topography of the ocean surface. In order to exploit this high-precision altimetry data, a goal of ~ 10 cm orbit accuracy has been set. With simultaneous estimation of all parameters (including GPS and Topex orbits), Topex orbit accuracy at the 5-10 cm level is expected [7]. Other future Earth orbiters which may carry GPS receivers include Gravity Probe-B, the Aristoteles gravity satellite, and the Earth Observing System (EOS) platforms. By the late 1990s, advanced GPS technology may enable user orbit accuracy at the level of a few cm [19]. High-Earth and elliptical orbiting satellites may also be able to utilize GPS flight receivers for precise orbit determination [20].

The *reduced-dynamic* GPS tracking strategy was designed to minimize the effect of gravity model uncertainty on the Topex orbit. By isolating the gravity accelerations, however, this same strategy can directly improve our knowledge of the gravity field [8] by as much as an order of magnitude for certain geopotential coefficients. For lower altitude satellites or for spacecraft undergoing maneuvers, less weight may be placed on dynamic models: in cases where the dynamics are unpredictable, the GPS data are relied on exclusively for the orbit solution. This is referred to as kinematic (or non-dynamic) tracking and will probably be the approach for precise orbit determination of large, low altitude platforms such as EOS and the Space Station [19].

Other GPS Applications

GPS signals can be used for precise determination of media propagation delays, in particular through the ionosphere [2] and through the troposphere [21]. Precise (sub-nanosec) clock sync or time transfer between distant tracking sites may also be feasible with GPS, which would be useful at astronomical observatories and NASA's deep space tracking sites. Analysis and experiments are underway at JPL to study the accuracy of clock sync with GPS.

FUTURE DEVELOPMENTS IN GPS

Receiver technology. Lighter, lower noise receivers are expected to be offered by a number of different manufacturers for high-precision applications. As production increases, the prices are expected to drop. A number of geodetic networks, including NASA's proposed FLINN (Fiducial Laboratories for an International Natural science Network), will deploy hundreds of GPS ground receivers for long-term geodetic and geophysical studies. GPS flight receivers should also improve in performance as well as in weight and power requirements during the 1990s as additional GPS instruments are flown in experiments which require higher and higher accuracy. A key issue for receiver designers is the handling of SA and AS and the precision of the pseudorange. Since policies may change for GPS transmissions with SA and AS, it is important that flexible codeless techniques be available if needed. As discussed above, there are different approaches to codeless receiver technology with significant implications for tracking performance.

Antenna technology. The minimization of antenna multipath is an important issue for GPS tracking applications. In some cases, multipath may be a limiting error source. GPS users will be looking for antenna designs with improved multipath suppression, particularly in difficult ground environments or for cluttered spacecraft where a clear view of the GPS constellation is not always available. The geometry of tracking GPS from satellites in high-Earth or elliptical orbits may require special antenna designs.

Algorithms and processing. Because of the relatively large number of GPS satellites (eventually to include 21 satellites plus three operational spares) and expected dozens of ground tracking sites, the efficiency of GPS data processing is a major aspect of GPS analysis. A continuously operating GPS geodetic network in California has just started to operate to routinely produce orbit and baseline solutions [22]. Other similar permanent networks exist in Canada and Japan. Streamlined data collection and analysis techniques are being utilized and automated procedures are used whenever possible, including autonomous operation of receivers. Communication between computers and receivers in the field, including downloading of data, occurs over phone lines. One of the key enhancements for such permanent, continuously operating GPS ground networks is deployment of an advanced GPS receiver which internally pre-processes the data, saving considerable computing time later. Ideally, such a receiver will contain a chip which is programmed to do many of the tasks, but operators will be able to communicate with the receiver remotely and control various parameters. A potentially fruitful area of research is parallel processing: it may be possible to streamline GPS data reduction by doing many computations in parallel before the final step of combination of data from different receivers. Included here would be filtering mechanizations which enable parallel methods to be used to maximal advantage. Incorporation of higher order ionospheric effects into the data processing and calibration software may be necessary for high-precision regional geodesy [13].

The 1980s saw the initial development of GPS positioning techniques with a relatively small subset of the GPS constellation, with demonstrated positioning accuracy of about 1 part in 10^8, corresponding to about 50-100 cm for GPS orbits, or about 1 cm for 1000 km ground baseline distance. The next few years will bring the filling out of the GPS constellation, the refinement of GPS technology, and the beginning of what will become routine and in some cases automated GPS tracking and data processing. In the 1990s, we expect to see further improvement to the level of a few parts in 10^9 (10-20 cm GPS orbit accuracy) and the first sub-decimeter orbits determined for Earth orbiters carrying GPS flight instruments.

Acknowledgment. The work described in this paper was carried out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.

REFERENCES

1. Thomas, J. B.: Functional Description of Signal Processing in the Rogue GPS Receiver. JPL Pub. 88-15, 1988.

2. Srinivasan, J. M.; Meehan, T. K.; and Young, L. E.: Code and Codeless Ionospheric Measurements with NASA's Rogue GPS Receiver. Proceedings, Inst. of Navigation GPS-89 Conference, 1989, pp. 451-454.

3. McCosh, D.: Navigating in Japan. Popular Science, Vol. 237, Oct. 1990, p. 30.

4. Lichten, S. M.: Estimation and Filtering for High-Precision GPS Applications. Manuscripta Geodaetica, Vol. 15, 1990, pp. 159-176.

5. Blewitt, G.: An Automatic Editing Algorithm for GPS Data. Geophys. Res. Lett., Vol. 17, 1990, pp. 199-202.

6. Blewitt, G.: Carrier Phase Ambiguity Resolution for the Global Positioning System Applied to Geodetic Baselines up to 2000 km. J. Geophys. Res., Vol. 94, 1989, pp. 10187-10203.

7. Wu, S. C.; Yunck, T. P.; and Hajj, G. A.: Towards Decimeter Topex Orbit Determination Using GPS. Paper AAS 89-359, AAS/AIAA Astrodynamics Specialist Conference, Aug. 7-10, 1989 (AAS Pub. Office, PO Box 28130, San Diego, CA).

8. Bertiger, W. I.; Wu, J. T.; and Wu, S. C.: Gravity Field Improvement Using GPS Data from Topex/Poseidon: A Covariance Analysis. Paper AIAA 90-2944, AIAA/AAS Astrodynamics Conference, Aug. 20-22, 1990 (AIAA, 370 L'Enfant Promenade, SW, Washington DC 20024).

9. Wu, J.T.; Bertiger, W. I.; and Wu, S. C.: Converting Gravity Bins to Spherical Harmonics. Paper AIAA 90-2943, AIAA/AAS Astrodynamics Conference, Aug. 20-22, 1990 (AIAA, 370 L'Enfant Promenade, SW, Washington DC 20024).

10. Freedman, A.: Measuring Earth Orientation with the Global Positioning System. Manuscripta Geodaetica, 1991 (in press)

11. Lichten, S. M.; and Neilan, R. E.: Global Networks for GPS Orbit Determination. Invited paper for Proc. of GPS'90—Second Int. Symp. on Precise Positioning with GPS, Ottawa, Canada, 1990 (in press).

12. Schutz, B. E.; Ho, C. S.; Abusali, P. A. M.; and Tapley, B. D.: CASA UNO GPS Orbit and Baseline Experiments. Geophys. Res. Lett., Vol. 17, 1990, pp. 643-646.

13. Blewitt, G.: GPS Techniques for Monitoring Geodynamics at Regional Scales. Invited paper for Proc. of GPS'90—Second Int. Symp. on Precise Positioning with GPS, Ottawa, Canada, 1990 (in press).

14. Shimada, S.; Fujinawa, Y.; Sekiguchi, S.; Ohmi, S.; Eguchi, T.; and Okada, Y.: Detection of a Volcanic Fracture Opening in Japan using GPS Measurements. Nature, Vol. 343, 1990, pp. 631-633.

15. Larson, K. M.: Precision, Accuracy, and Tectonics from the Global Positioning System. Doctoral Thesis, Univ. of California, San Diego, 1990.

16. Lindqwister, U. J.; Lichten, S. M.; and Blewitt, G.: Precise Regional Baseline Estimation Using A Priori Orbital Information. Geophys. Res. Lett., Vol. 17, 1990, pp. 219-222.

17. Lichten, S. M.: Towards GPS Orbit Accuracy of Tens of Centimeters. Geophys. Res. Lett., Vol. 17, 1990, pp. 215-218.

18. Ruland, R.; and Leick, A.: Application of GPS in a High Precision Engineering Survey Network. Proc. of First Int. Symp. on Precise Positioning with GPS, Vol. I, 1985, (U.S. Dept. of Commerce) pp. 483-494.

19. Yunck, T. P.; Wu, S. C.; Lichten, S. M.; Bertiger, W. I.; Lindqwister, U. J.; and Blewitt, G.: Toward Centimeter Orbit Determination and Millimeter Geodesy with GPS. Proc. of 5th Int. Geodetic Symp. on Satellite Positioning, Vol. I, 1989 (Physical Science Lab., New Mexico State Univ.) pp. 272-281.

20. Lichten, S. M.; and Estefan, J.: High-Precision Orbit Determination for High-Earth Elliptical Orbiters Using the GPS. Paper AIAA 90-2954, AIAA/AAS Astrodynamics Conference, Aug. 20-22, 1990 (AIAA, 370 L'Enfant Promenade, SW, Washington DC, 20024).

21. Lichten, S. M.: Precise Estimation of Tropospheric Path Delays with GPS Techniques. TDA Progress Report , 42-100, Feb. 15, 1990, pp. 1-12.

22. Lindqwister, U. J.; Blewitt, G.; and Yunck, T. P.: Continuous Monitoring of GPS Geodetic Networks. Proc. of GPS'90—Second Int. Symp. on Precise Positioning with GPS, Ottawa, Canada, 1990 (in press).

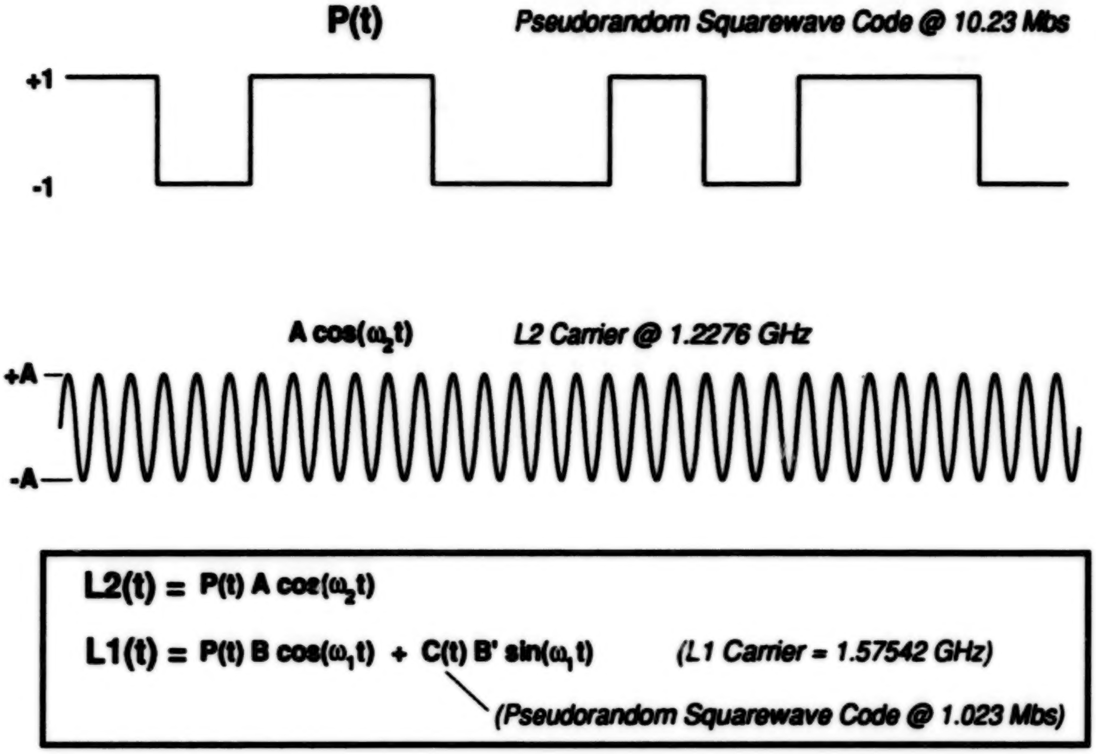

P(t)
Pseudorandom Squarewave Code @ 10.23 Mbs
+1
-1
A cos(ω2t)
L2 Carrier @ 1.2276 GHz
+A
-A
L2(t) = P(t) A cos(ω2t)
L1(t) = P(t) B cos(ω1t) + C(t) B' sin(ω1t)
(L1 Carrier = 1.57542 GHz)
(Pseudorandom Squarewave Code @ 1.023 Mbs)

Fig. 1 GPS carrier phase is modulated by the P-code [P(t)] and the C/A code [C(t)]. The codes include information about the GPS orbits and clocks. Carrier phase and high-precision pseudorange obtained from either code or codeless tracking techniques are used in high-accuracy GPS applications. For relatively coarse accuracy applications, an inexpensive C/A-code only GPS receiver may be adequate.

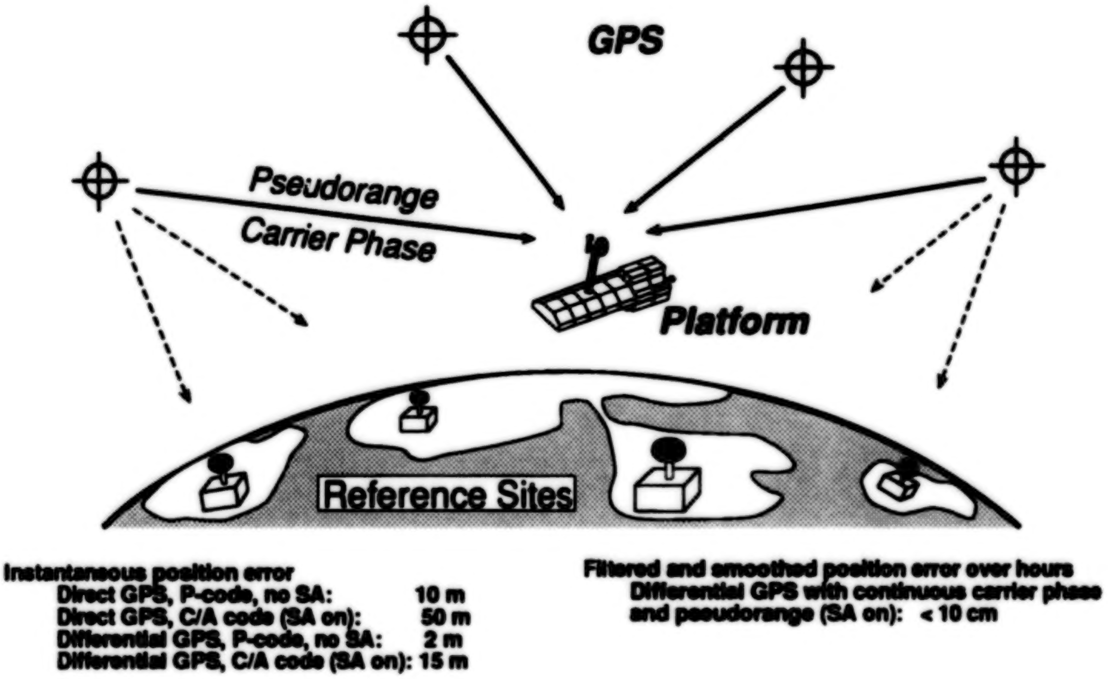

GPS
Pseudorange
Carrier Phase
Platform
Reference Sites
Instantaneous position error
Direct GPS, P-code, no SA: 10 m
Direct GPS, C/A code (SA on): 50 m
Differential GPS, P-code, no SA: 2 m
Differential GPS, C/A code (SA on): 15 m
Filtered and smoothed position error over hours
Differential GPS with continuous carrier phase and pseudorange (SA on): < 10 cm

Fig. 2 Simultaneous GPS tracking data from ground and/or orbiting receivers can be combined to eliminate clock and clock-like errors and substantially reduce the effect of GPS orbit error. Non-differential (direct) instantaneous positioning accuracy is generally much less accurate. For very high accuracy applications, carrier phase continuously tracked for hours and processed together with pseudorange should enable sub-10 cm positioning accuracy in non-real time.

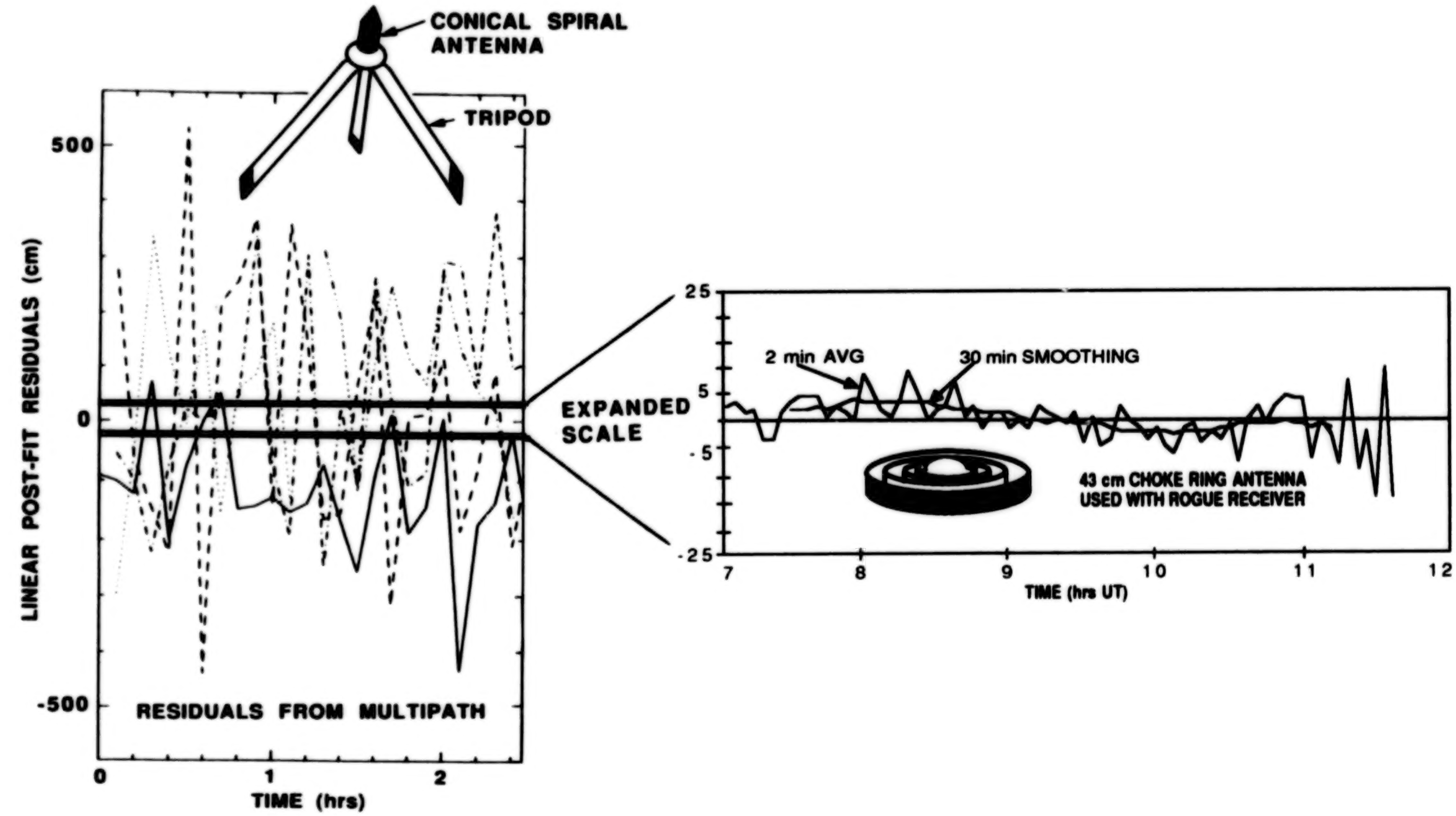


Fig. 3 A number of commercially available GPS instruments use an antenna similar to the one shown at the left. This configuration produces high levels of ground multipath for GPS pseudorange, resulting in high measurement scatter. The choke ring antenna on the right reduces the multipath by more than an order of magnitude.

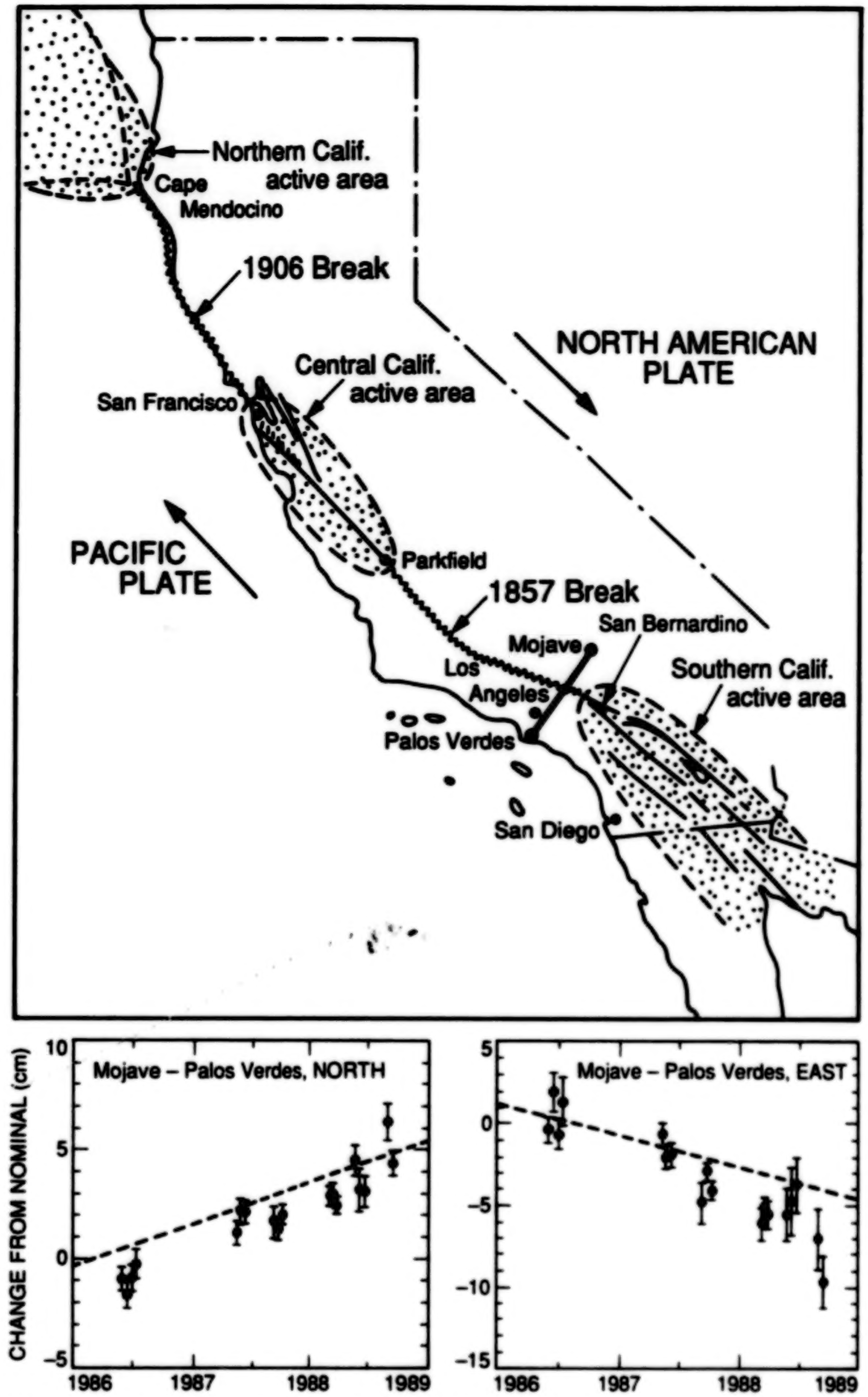

Northern Calif. active area
Cape Mendocino
1906 Break
NORTH AMERICAN PLATE
Central Calif. active area
San Francisco
PACIFIC PLATE
Parkfield
1857 Break
San Bernardino
Mojave
Los Angeles
Southern Calif. active area
Palos Verdes
San Diego
Mojave – Palos Verdes, NORTH
Mojave – Palos Verdes, EAST
CHANGE FROM NOMINAL (cm)
10
5
0
−5
−10
−15
1986
1987
1988
1989

Fig. 4 Map of California shows the San Andreas fault where the North American and Pacific plates meet and are moving relative to one another. Large earthquakes in 1857 and 1906 occurred along the fault as shown. Dozens of GPS instruments have been monitoring the crustal motion in Southern California over the past several years, including GPS receivers at Mojave and Palos Verdes, which form a baseline crossing the San Andreas fault. Bottom graphs from [15] show the plate motion in measured with GPS: the Pacific plate is moving to the northwest at ~ 5 cm/yr. Dashed line shows the rate as measured independently by VLBI (radio astrometric) techniques: the VLBI and GPS rates agree to a few mm/yr.

SESSION G - SENSORS AND MEASUREMENT TECHNOLOGY (PART 2)

Wednesday November 28, 1990

- **Quantitative Nondestructive Evaluation - Requirements For Tomorrow's Reliability**
- **Robotic Control And Inspection Verification**
- **The Transfer Of A Technology To Measure Skin Burn Depth In Humans**
- **Frequency Domain Laser Velocimeter Signal Processor**
- **A Field-Deployable Digital Acoustic Measurement System**
- **Laser-Optical Disk Position Encoder With Active Heads**

N91

24074

UNCLAS

N91-24074

QUANTITATIVE NONDESTRUCTIVE EVALUATION - REQUIREMENTS FOR TOMORROW'S RELIABILITY

Joseph S. Heyman
NASA Langley Research Center
Hampton, Virginia 23665-5225

ABSTRACT

Quantitative Nondestructive Evaluation (QNDE) is the technology of measurement, analysis, and prediction of the state of material/structural systems for safety, reliability and mission assurance. QNDE has impact on everyday life from the cars we drive, the planes we fly, the buildings we live/work in, literally to the infrastructure of our world. In this paper, we will highlight some of the new sciences and technologies that are part of a safer, cost-effective tommorow. Specific technologies to be discussed are: thermal QNDE to aircraft structural integrity, ultrasonic QNDE to materials characterization, and technology spin-off from aerospace to the medical sector. In each case, examples will be given of how new requirements result in enabling measurement technologies, which in turn, change the boundaries of design/practice.

INTRODUCTION

In our ever increasingly complex world, we must rely on the integrity of our surroundings for safety, reliability and economy. That is equally true for ballpoint pens, computers, cars, power plants, or hypersonic vehicles. Although all premature failures are undesirable, the associated risk can range from annoyance to catastrophe. A critical vitalized technology, quantitative nondestructive evaluation (QNDE), is bringing new reliability to industry starting with design concepts, prototype verification, process control, end point inspection, and in-service inspection leading to maintenance repair, or retirement-for-cause. In this paper, several recent NASA developments are highlighted, which are ready for technology transfer and application.

Design takes advantage of QNDE in the fine tradition of concurrent engineering. Using CAD (computer-aided design) tools, a proposed design can be evaluated for inspectability based on realistic QNDE techniques: for example, ultrasonics, magnetics, thermal, or radiography. In addition, a probability assessment for missing a flaw can be established and used for fracture mechanics predictions of safe-life. Therefore, before metal is cut (or polymers are cured), one can start with a basis of design reliability.

A second stage in most product development is prototyping. Again, QNDE plays a critical role helping to characterize fabrication problems, stress profiles, design shortcomings, and degradation occurring in accelerated life testing.

The third stage is process monitoring, which is a key cost-effective approach to catching problems before and during fabrication by applying feedback to the process thereby raising overall quality. For example, ultrasonics can monitor the degree of cure and viscosity of polymer systems during manufacturing.

After fabrication, for critical structural elements, it is often desirable to verify integrity. Such verification may require geometry or material property tests. Again, QNDE can check detailed geometry with data from a three-dimensional x-ray CAT (computer axial tomography) scanner compared to CAD design drawings. Material properties can be assured through physical measurements of stiffness (using ultrasonics and x-ray), density (CAT), and fiber orientation (ultrasonics or thermography). Flaws can be found, sized, located, and assessed using structural mechanics to predict the impact of such findings on performance.

Finally, life cycle costs can be reduced while safety and reliability is improved if QNDE is applied during appropriate intervals throughout the element's service life.

QUANTITATIVE NDE TECHNOLOGIES DEVELOPMENT/TRANSFER

In this paper, several recent developments in QNDE are highlighted. In each case, these developments resulted from well defined needs. Typically, physical models of the proposed concept were developed to guide and focus the experiments needed to verify the approach. The models permit optimization of approaches before major investments are required for experimental hardware. During the evolution of these concepts, there is a constant interplay between experimental data and model predictions--moving step-by-step through the process of development. If tests are successful, a prototype instrument is built and tested on real (or realistic) hardware.

Technology developed in this "model first" fashion holds great promise for transfer to industry. First, the data generated by this process is quantitative, science based and therefore more readily interpreted through the model. The market potential is usually broad since the solution tends to be generic--that is, it can be applied to other similar problems by changing elements in the model to identify how to modify the instrument. In contrast, correlative development approaches tend to work only on the problem at hand and require extensive testing if any of the variables are changed. The technologies highlighted here have been brought through the outlined process and as such represent reduced risk opportunities for U.S. industry. The successful transfer is of benefit to NASA as well as to the company involved. New QNDE developed by NASA is of maximum value when it is commercially available to NASA and its contractors through a company source. The applications of these technologies is usually to a broad sector of which aerospace is but a small part. Thus, there is benefit to all parties involved.

SPECIFIC TECHNOLOGIES - THERMAL QNDE

Thermal QNDE has many benefits for inspection [1,2]. It can be noncontacting, view large areas at a single time, and result in cost-effective measurements for practical structures. The method finds flaws through their effect on heat flow in objects. In one rendition of the technique, some form of heat energy is injected into a surface--optically, electrically, with hot water or with some other method. Only a few degrees temperature difference is required. The infrared (IR) radiation emitted from the heated surface as a function of time is monitored with an IR camera and its image is stored. The acquired data is processed based on a physical model of the tested object to interpret the temperature data to reveal hidden internal structural information. Techniques have been developed to minimize spurious effects such as background reflections, nonuniform heating and surface emissivity variations (related to the efficiency of a surface to emit radiation).

A typical measurement setup is shown in figure 1. In this situation, quartz lamps are used as a heat source and a digital image processor is used to acquire and process the data. The system is under computer control such that the heat input/data acquisition has been optimized based on simulation analysis using material properties and geometry of the tested structure.

An example of the utility of this approach is shown in figure 2. The image is of a flat aluminum plate of 0.040" thickness bonded to a second plate that has been cut to form letters spelling out "NASA LaRC NDE." The letters simulate hidden damage as might occur from fatigue or corrosion.

Figure 2a is a normal photograph of the plate while figure 2b is an image obtained from LaRC's thermal QNDE system clearly "seeing" through the plate to the hidden back surface letters.

Figure 3 demonstrates a practical application of this system for imaging the internal structure of an airplane fuselage. Figure 3a is a mechanical drawing of the internal fuselage showing the waffle-like design while figure 3b shows the ability of the QNDE approach to help assess the integrity of structural elements by "seeing" through the outer skin to the hidden internal structure.

LaRC has developed and applied thermal QNDE approaches to a wide array of problems ranging from diamond film coatings on electronic materials to solid rocket motor insulation adhesive bonds. In each case, the problem is first analyzed, then modeled, and a testing approach developed/optimized. Many problems are so similar that test protocols can be developed by a simple modification of previous approaches. This technology has high promise for development, sales, and service.

SPECIFIC TECHNOLOGIES - BOLT TENSION MONITOR

The lowly bolt may seem to be a strange component for advanced high technology--until one realizes its importance in many critical applications. From bridges to aircraft, integrated power circuits to power plants, bolts are transformed from crude mechanical devices to precise "instruments" required to do their job. In many situations, a bolt failure can have catastrophic consequences. The difficulty in tensioning fasteners is that one cannot easily and directly measure the clamping force exerted by a bolt. Instead, that value is inferred from other measurements such as torque. Torque in particular does not measure what a bolt does: it only measures "twist." Frictional variations make torque an inaccurate tensioning method resulting in errors that can be appreciable [3]. Shown in figure 4, is a plot of a space-quality fastener tensioned 20 times to a desired load of 46,000 pounds force. Note that the torque to achieve that load varies by over 200 percent during the tensioning sequences!

To overcome these problems, NASA LaRC has developed a bolt tensioning monitor that more directly measures the force in the bolt [4]. The technique is based on ultrasonic propagation inside the bolt. In contrast to some ultrasonic systems which measure time, this device measures the acoustic phase shift which accompanies bolt tension as the bolt gets longer (similar to a rubber band, a bolt doesn't perform its function unless it is stretched). Figure 5 shows the superior accuracy of ultrasonics over torque for the tensioning of fasteners using different lubricants to vary friction. Whereas the torque curves vary dramatically for each lubricant tested, the ultrasonic curves are nearly identical for all tests.

The LaRC system, shown in figure 6 is a computer-based ultrasonic instrument ready for field applications. Important data is taken during bolt clamp-up and is archived in the device's memory. After the day, the memory card can be removed and the records of all fasteners tensioned can be stored for future reference to insure integrity. Furthermore, at some later time, the fastener tension can be recertified to insure it has not relaxed in-service. The technique is fairly forgiving of bolt geometry since it uses a frequency wave which propagates more "well behaved" than a pulse which contains all frequencies. This system is at the prototype stage and will be transferred to industry through one of our technology transfer mechanisms such as licensing or a joint development agreement.

SPECIFIC TECHNOLOGIES - MEDICAL TECHNOLOGY TRANSFER

NASA scientists and engineers are made aware of critical non-aerospace problems affecting other sectors of the national infrastructure to maximize Agency technology transfer opportunities. One mechanism for that awareness is the Technology Utilization Program which circulates listings of problems brought to NASA's attention by other agencies, industry, and universities. Often, when an expert in another field identifies a problem, it is identified as a request for a specific measurement rather than a well defined problem statement. Working with the individual, however, the measurement science people help identify the nature of the problem and thus focus on a direct path to its solution.

Bladder incontenence is an example of a medical problem brought to our attention that afflicts millions of Americans today. The problem was identified as a requirement for measuring the diameter of the bladder to determine its fullness to alert a wearer of the need for elimination. In looking at the problem, it was determined that the stated solution based on bladder diameter was already in use, but additional unstated complexity compromised its success. The problem needed to be approached differently.

The bladder is not a simple inflating sphere. Therefore, a measure of the front to back surface diameter is not a reliable indication of fullness. Instead, the bladder is more like a slightly elastic pouch whose geometry differs from person to person. One of the major changes that occurs during distension, is the tension in the bladder material itself. It varies from a flaccid film to a stretched membrane. Ultrasonics, incident on the bladder, is transmitted to back surface tissue easily when the bladder is full. When empty, the bladder wall scatters/attenuates acoustic energy resulting in a reduction in the reflection/reverberation at the back wall. This, in turn, reduces the measured signal amplitude which is identified as an empty bladder.

Figure 7 shows a block diagram of the bladder monitor developed for this program [5]. It is a self-contained ultrasonic system. It launches an ultrasonic wave into the bladder, digitizes the returning signals, stores the waveform in multichannels, compares the signals with stored patient training signals, and, if necessary, alerts the wearer to bladder fullness. In actual practice, the device spends most of its time "asleep" extending its battery life, waking up to take measurements on a periodic basis.

The wearer can be alerted by several modes of communication: audio, visual, tactile, and remote. When triggered, the audio device sounds a small alarm, the visual device lights a small LED in the wearer's glasses, the tactile device taps the skin, and the remote device sends a radio wave to a separate station (operated by a care-giver) and sounds an alarm remotely.

Figure 8 shows the bladder distension device concept interrogating a nearly empty bladder (top) and a full bladder (bottom). As the back wall undergoes tension, it results in an increase in reverberation which is detected in the ultrasonic response. The monitor takes several features into account in alarming for a full bladder. The device determines a scan value which is a function of the distance from the bladder front surface to the back surface multiplied by a weighing constant modified by the energy return in the reverberation signal. The scan value is computed for each wearer as an alarm threshold thereby personalizing the operation of the system. In operation, the threshold must be achieved several times in succession to indicate bladder fullness. Such redundnacy reduces the potential of false calls.

In another medical technology transfer, a problem was identified requiring measurement of internal body temperature during hyperthermia treatment of cancer patients. The current practice involves multiple invasion of the region with thermocouples, often repeatedly, to determine the temperature profile around the tumor. Instead, LaRC suggested using ultrasonics to monitor the phase transformation of a wax material injected near the tumor site [6]. The wax can be prepared to melt at the desired treatment temperature and to remain at the tumor site during the entire treatment period. As the wax melts, its acoustic impedance changes which can be determined noninvasively as a large ultrasonic signal fluctuation indicating that the heating has reached the desired treatment level.

Figure 9 shows a concept drawing of the temperature monitoring device with data indicating the effect of heating the tumor to the wax melting point. The top part of the figure shows the ultrasonic data for unmelted wax. The high difference in impedance between the wax and the surrounding tissue reduces the amplitude of the signal reflecting off of the tumor. In contrast, after the wax melts, as is shown in the lower part of the figure, the reflection off of the tumor is large. Therefore, the proper temperature for treatment can be maintained by observing the state of the wax and controlling the heating power accordingly.

SUMMARY

These are but a few of the many opportunities in the field of quantitative nondestructive evaluation for technology transfer through one or more of the many mechanisms available through NASA. Such transfer is of value to the receiving industry tapping the intellectual resources of our National labs. In addition, the transfer is of value to the Agency bringing needed technology to commercialization so that it can be widely utilized by NASA and its contractors. In addition, the access to state-of-the-science QNDE technology brings value to the public sector through improved safety, reliability, and cost-effectiveness in addition to the synergism of accelerating applications for non-aerospace industries.

ACKNOWLEDGEMENTS

The author thanks many of the staff of the Nondestructive Measurement Science Branch, NASA LaRC for their contributions to this paper and to their research applications which it represents. In particular, I appreciate the help of Scott Crews, Sid Allison, and John Companion for their assistance in preparing figures for the paper.

REFERENCES

1.) Welch, Christopher S. ; Winfree, William P. ; Heath, D. Michele ; Cramer, Elliott and Howell, Patricia : Material Property Measurement With Post-Processed Thermal Image Data. SPIE, 1313, pp. 124-133, 1990.

2.) Winfree, William P. and James, Patricia H.: Thermographic Detection Of Disbonds. In Proc. of 35th International Instrumentation Symposium (ISA), pp. 183-188, 1989.

3.) Heyman, Joseph S. and Chern, E.J. : Ultrasonic Measurement Of Axial Stress. Jour. Testing and Evaluation, 10, pp. 202-211, 1982.

4.) Allison, S.G. and Clendenin, C. G.: Ultrasonic Pulsed Phased Locked Loop Interferometer For Bolt Load Measurements. In Proc. of 35th International Instrumentation Symposium (ISA), pp. 189-194, 1989.

5.) Companion, J.; Heyman, Joseph S.; Blalock, T.; Cavalier, A.; Mineo, B.; Klien, F.; and Fox, L. : Bladder Distension Sensor For The Handicapped. In Proc. IEEE Ultrasonics Symposium, pp. 953-956, 1986.

6.) Companion, J. ; Heyman, Joseph S. ; Oleson, James R.; Engler, Mark: Ultrasonic Hyperthermia Temperature Monitoring Technique. In Proc. IEEE Ultrasonics Symposium, pp. 977-980, 1986.

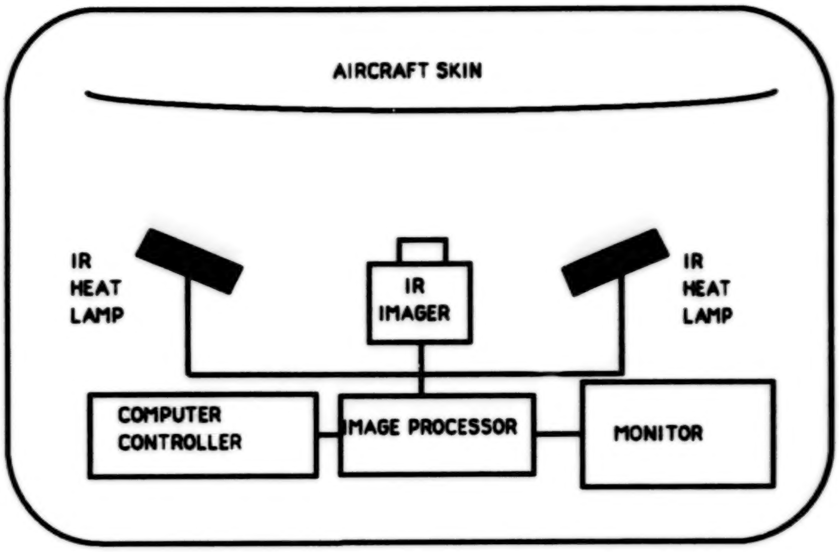

AIRCRAFT SKIN
IR HEAT LAMP
IR IMAGER
IR HEAT LAMP
COMPUTER CONTROLLER
IMAGE PROCESSOR
MONITOR

Figure 1: Thermography NDE Measurement System

ORIGINAL PAGE IS
OF POOR QUALITY

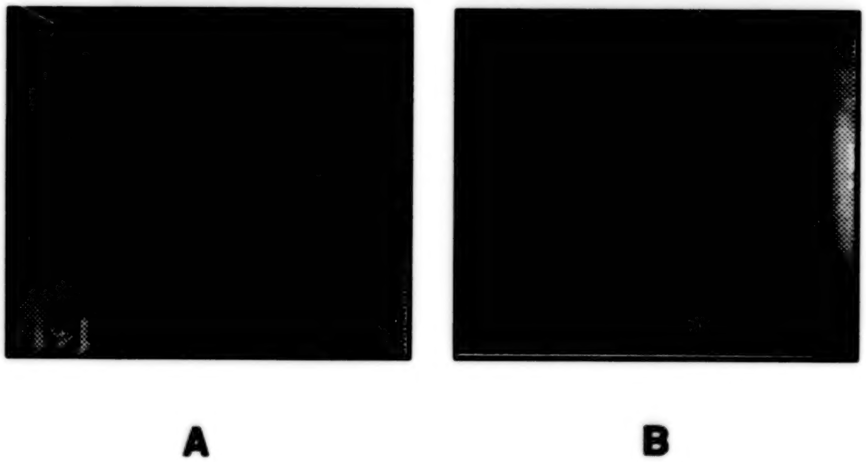


Figure 2: Front Surface of a Test Panel.

(A) Normal Video Camera

(B) LaRC Quantitive Thermal NDE Measurement System

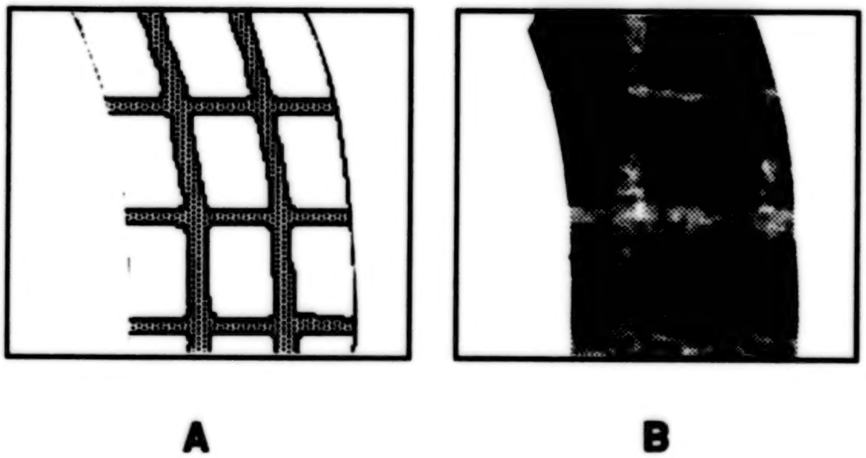


Figure 3: Aircraft Fuselage Subsurface Structure.

(A) Shaded Areas are Structural Strengtheners

(B) Doublers as Imaged by QNDE System for Integrity Assesment

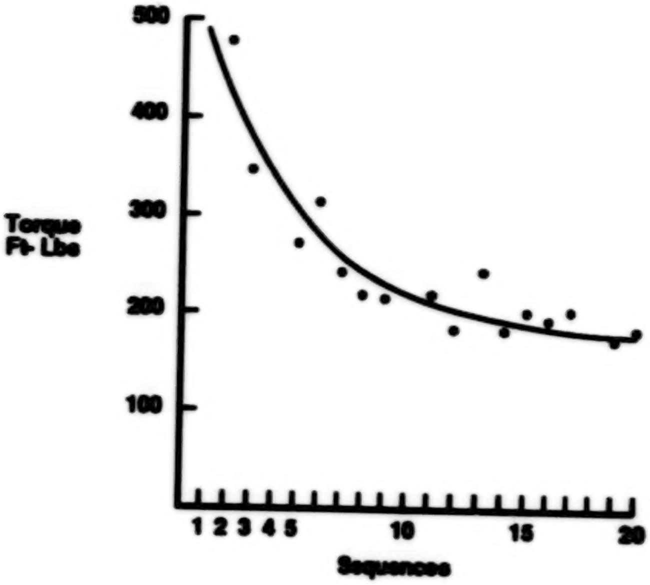


Figure 4: Effect of Multiple Tensioning Sequences on Friction

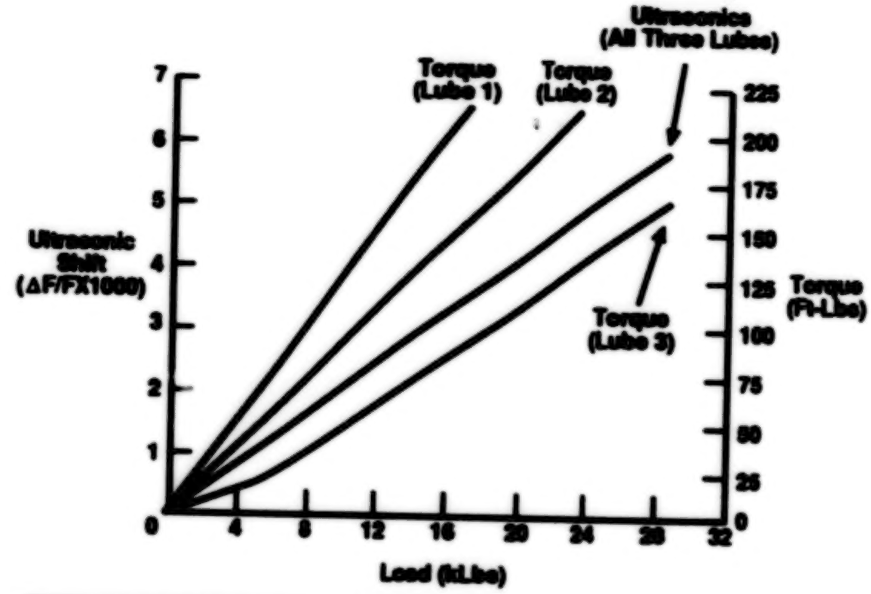


Figure 5: Comparison of Tensioning with Torque Versus Ultrasonics for Different Lubricants

ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6: Field Portable Bolt Monitor Shown with Testing Apparatus

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

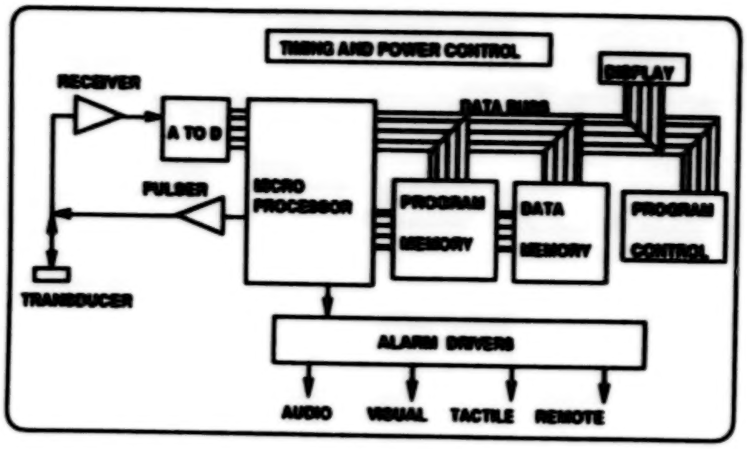

TIMING AND POWER CONTROL
DISPLAY
RECEIVER
A TO D
DATA BUSS
PULSER
MICRO PROCESSOR
PROGRAM MEMORY
DATA MEMORY
PROGRAM CONTROL
TRANSDUCER
ALARM DRIVERS
AUDIO
VISUAL
TACTILE
REMOTE

FIGURE 7 ULTRASONIC BLADDER DISTENSION MONITOR BLOCK DIAGRAM

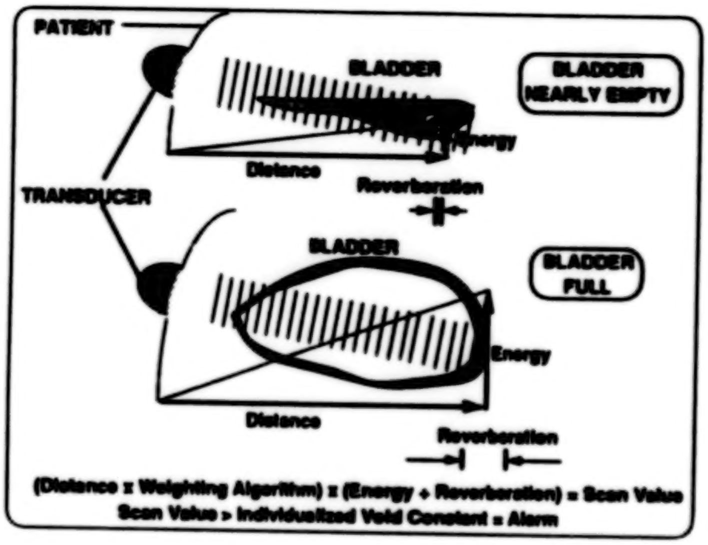

PATIENT
BLADDER
BLADDER NEARLY EMPTY
Energy
Distance
TRANSDUCER
Reverberation
BLADDER
BLADDER FULL
Energy
Distance
Reverberation
(Distance x Weighting Algorithm) x (Energy + Reverberation) = Scan Value
Scan Value > Individualized Void Constant = Alarm

FIGURE 8 ULTRASONIC BLADDER DISTENSION MONITOR CONCEPTIONAL DIAGRAM

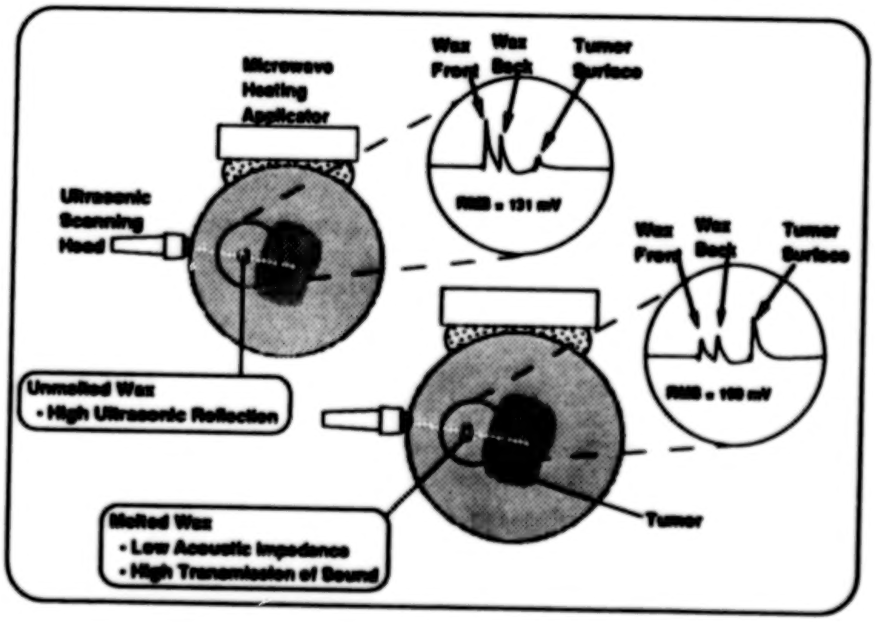

Microwave Heating Applicator
Wax Front
Wax Back
Tumor Surface
Ultrasonic Scanning Head
RMS = 131 mV
Wax Front
Wax Back
Tumor Surface
RMS = 109 mV
Unmelted Wax
• High Ultrasonic Reflection
Melted Wax
• Low Acoustic Impedance
• High Transmission of Sound
Tumor

FIGURE 9 ULTRASONIC TUMOR TEMPERATURE MONITOR CONCEPTIONAL DIAGRAM

ORIGINAL PAGE IS
OF POOR QUALITY

N91-24075

ROBOTIC CONTROL AND INSPECTION VERIFICATION

Virgil Leon Davis
Chief, Robotics Section, DM-MED-12
Kennedy Space Center, FL

ABSTRACT

This paper discusses three possible areas of commercialization involving Robots developed at John F. Kennedy Space Center (KSC), Florida.

(1) A 6-Degree of Freedom (6-DOF) Target Tracking System for remote umbilical operations

(2) An intelligent Torque Sensing End-Effector (TSEE) for operating hand valves in hazardous locations

(3) An Automatic Radiator Inspection Device (ARID), a 65 by 13 foot robotic mechanism involving completely redundant motors, drives, and controls

Aspects concerning the first two innovations can be integrated to enable robots or teleoperators to perform tasks involving orientation and panel actuation operations that can be done with existing technology rather than waiting for telerobots to incorporate artificial intelligence (AI) to perform "smart" autonomous operations. These operations are applicable to Space Station work, ground aerospace launch processing, and hazardous petrochemical or nuclear safing operations worldwide. The third robot involves the application of complete control hardware redundancy to enable performance of work over and near expensive Shuttle hardware. The consumer marketplace may wish to explore commercialization of similar component redundancy techniques for applications when a robot would not normally be used because of "reliability" (when an inadvertent move could result in damage to expensive components or personnel).

Introduction

The computer hardware and software systems in the Robotic Applications Development Laboratory (RADL) were designed to facilitate the development and application of advanced robotic control technology. KSC not only launches spacecraft but services these spacecraft on the ground: designing the support equipment, launch accessories, and computer hardware/software for ground spacecraft servicing.

KSC has implemented an integrated system that coordinates state-of-the-art robotic subsystems. It is a sensor-based realtime robotic control system performing operations beyond the capability of an off-the-shelf robot. The integrated system provides realtime closed-loop adaptive path control of position and orientation of all six axes of a large robot; enables the implementation of a highly configurable, expandable testbed for sensor system development; and makes several smart distributed control subsystems (robot arm controller, process controller, graphics display, and vision tracking) appear as intelligent peripherals to a supervisory computer that coordinates the overall system.

The integrated RADL system is currently providing an easy-to-use testbed for NASA sensor integration experiments. Advanced target tracking development is in progress concerning the mating of umbilicals used during space vehicle launch. Programmatic studies are underway to use laboratory capabilities to enhance the safety, productivity, and efficiency of KSC facilities for Shuttle and future ground processing operations. Projects are underway that should generate large operational cost savings through the integration of advanced technologies for ground processing operations, such as Orbiter tile and radiator damage assessment.

Robotic techniques to improve "Shuttle Orbiter inspection and closeout verification" (operations involving possible human- or mechanism-induced damage) are being investigated and implemented. Nondestructive test sensors, vision systems, and various kinds of distance ranging sensor systems can be integrated with the RADL systems to develop the prototype concepts for integrating robot parameters with large data-based graphics and artificial intelligence software systems. For example, the RADL robot can position a sensor with precise accuracy, report that position and orientation, provide distance sensory data, and integrate machine vision "electronic photographs" with graphics and AI software to furnish computer printouts providing automatic sizing and highlighting of exception data. This type of system is being proposed for Shuttle Orbiter radiator damage inspection, Orbiter tile damage/debonding assessment, and Orbiter contour measurements. The manual methods presently employed in these operations are very labor intensive and produce expensive serial-time flow constraints.

REMOTE UMBILICAL PLATE DOCKING/INSERTION

Realtime adaptive control is the necessary tool for tracking a Shuttle vehicle that rocks in the wind while stacked at the launch pad. This adaptive control is necessary in order to dock and insert umbilicals (consisting of a ganged connection of electrical and cryogenic/hypergolic fluid lines) without damage to the vehicle and without hazardous leaks. This reduces reconnect times of 14 to 34 hours to less than 15 minutes and eliminates hazards associated with umbilicals that would otherwise have to be connected at launch. The KSC environment is demanding: the system must withstand heavy acoustical shock and see through fog caused by the dumping of thousands of gallons of water onto the flames of the Shuttle main engines (conditions that exist right after an aborted launch when an umbilical reconnect would take place). These shock and blast conditions rule out sensitive laser tracking. The prototype system KSC is developing can be upgraded quite easily for this environment. For instance, an infrared filter can be added to the CCD camera and our 5 dots can be changed to infrared LED's to see through the water vapors without affecting the system architecture or changing the extensive software algorithms.

Also, KSC is tracking speeds at which the target can be blurred. KSC has already advanced the state of the art by developing algorithms and packaging off-the-shelf imaging hardware into process machines that significantly reduce the blurring of moving targets to allow more precise, smoother tracking.

KSC is studying adaptive control (trajectory perturbation based on realtime sensory feedback) with heavy inertial loads. KSC is attempting to precisely position 5,000-pound umbilicals (200 pounds initially) with pneumatic counterbalances on future heavier umbilicals. Research in industry has been done on compliance-aided insertion of resisters into mm-tolerance holes; however, KSC is designing heavy umbilicals (oriented perpendicularly with respect to gravity) to insert into mm-tolerance holes on an object randomly moving with wind-induced perturbations. This induces high torques into servo control motors and KSC has obtained stability under these extreme conditions.

The RADL is located in a 50 by 100 foot high bay facility. Figure 1 depicts the left half of the facility. It has been recently expanded to provide space for a production model of the ARID robot. Several work cells (as shown in figure 1) are accessed by an ASEA IRB-90 six-axis industrial robot located on a 30-foot track. The IRB-90 has a reach of approximately 9 feet, a load capacity of 90 kilograms, and a repeatability of 0.005 inch. A central MicroVAX II mini-computer acts as the supervisory controller. Communication to peripheral systems (ASEA robot controller, AD/DA interface, and vision system) is established through custom serial connections. The system is connected via DecNet network to the local KSDN network and the National NASA SPAN network. The vision system, a Data Cube MaxVideo, incorporates a pipeline design approach, a VME bus with a 68020 processor, and a processed throughput of 30 frames per second. Additional sensors (force/torque, proximity) are also integrated in the system. Because of associated limitations, the original robot controller is now being replaced by a special-purpose high-speed flexible robot controller, which will provide direct access to each joint controller and provide 6-DOF adaptive control capabilities in realtime (30 Hz control updates for vision and 100 Hz for force-torque tracking).

The RADL prototype system consists of a vision-based 6-DOF tracking system attached to the ASEA robot and a target attached to a separate receptacle plate. Using a passive compliance end-effector, the robot is able to track and insert an umbilical plate mockup (which incorporates fluid, electrical, and data lines) into a receptacle plate mounted on a 3-DOF device that simulates the motion of the Orbiter (see figure 2). The passive compliance device uses nontactile vision tracking and is being augmented by an active tactile tracking system using force/torque reactive feedback to reduce mating forces.

Enhancements include the use of a counterbalance mechanism that removes loads from the robot and enables the robot to disconnect from the umbilical after mating. The robot will be turned off and the umbilical will free float with the counterbalance removing loads from the vehicle. This reduces vehicle design weight and allows more payload to reach space at reduced costs. A floating plate is also incorporated that reduces insertion loads, reduces connect/disconnect surges, and eliminates galling of sensitive cryogenic fluid coupling surfaces. Improvements also have been made in alignment and capture mechanisms to reduce forces, reduce galling, and ensure positive latching.

6-DOF Tracking

The major innovation is a robotic vision subsystem that measures the relative position and orientation of a specially designed target and provides realtime control to a large robotic mechanism. The subsystem uses standard image processing algorithms implemented directly in circuitry instead of computer programs that consume more time. This feature makes it possible to extract complete sets of target tracking data from successive image frames at the rate of more than 30 frames per second.

A solid-state video camera views the target, which consists of five bright or reflective circles, four located at the corners of a square and the fifth located at the center of the square but offset from the plane of the square (see figure 3). The raw image data is sent to image processing circuitry that performs a convolution difference-of-Gaussian edge-analysis filtering operation to clarify the picture elements representing the edges of the circles. On the Shuttle, the fifth circle could be implemented simply as a styrofoam thread-spool glued in the center of four painted circles.

The image is then processed further to obtain the centroids of the five circles. The locations of these centroids relative to each other and to the overall image frame are processed to obtain three Cartesian coordinates of the target relative to those of the camera. Triangulation calculations based on the vector relationships among the locations of the five circles and the central axis of the target yield the roll, pitch, and yaw angles that describe the orientation of the target relative to the line of sight and the field of view of the camera (see figure 4). Thus, the relative position and orientation of the target are determined in all six degrees of freedom (see figure 5). The offset of the central circle from the plane of the other four circles can be increased or decreased to increase or decrease the sensitivity of the subsystem to the pitch and yaw of the target and to provide more accurate distance-to-the-target information.

The output data may have to be transformed into spherical or other coordinates used by any other robot. However, this transformation can be performed easily in software. If the robot is changed, it is necessary only to change this software.

TORQUE SENSING END-EFFECTOR (TSEE)

The TSEE was produced from a Small Business Innovation Research (SBIR) project to develop an intelligent tool/gripper to open or close valves during hazardous maintenance or emergency work. The TSEE features servo control of jaw opening dimensions, nontactile/tactile sensors, and torque feedback to determine and maintain optimum seat pressure settings. This feedback provides reflective force feedback to an operator and to automatic computer-controlled operations enabling determination of valve position and preventing damage to valve seals. A Phase I SBIR (see figure 6) produced a small working model and a Phase II SBIR (see figure 7) produced a hardened mechanism with a user-friendly data base capable of operating in hazardous NEC Class I, Division I, Group B hypergolic/cryogenic environments. A version closer to the Phase I unit may be more suitable for petrochemical and nuclear applications where a smaller mobile robot (see figure 8) could be used to safe hazardous fires, or chemical or nuclear spills.

The TSEE's interchangeable gripper can locate a valve position and rotate continuously to open or close valves ranging from either 0.5 to 4.0 inches or 3.5 to 6.0 inches in size. Valves can be opened that have a torque range of 10 to 150 inch-pounds.. The TSEE has a unique nontactile torque sensor utilizing magnetoelastic phenomena.

The end-effector is used in conjunction with a computer controller that can interpret commands from an operator at a computer keyboard, from a parallel digital interface on the robot carrying the end-effector, or from a serial communications link to another computer. This lets the system open and close valves and determine if the valve is turning as expected from any of these operating modes.

The computer that controls the end-effector stores information about a range of valves whose positions need to be adjusted. Valves are identified to the system by labels that may be descriptive or numeric. Information in the valve data base is used to identify valves and to provide information on turning ranges, gripping forces, valve handle sizes, and current status or position of the valve. Once information has been put into the system, valve operations will keep the data base up to date.

A TSEE can be incorporated with the 6-DOF target tracking system to allow a remote operator (Space Station astronaut, Flight Telerobotic Servicer (FTS) remote controller, nuclear cleanup engineer, or fire fighter) to: (1) position or teleoperate the robot in front of a panel where it can see the target, (2) let the target tracker autonomously orient the robot at the correct angle and distance from the panel, and (3) activate a stored program to perform pretaught routines. The TSEE locates valves (which may be in the wrong position with stems in or out) through simple sensors and then performs pretaught operational sequences to "safe" an operation in a hazardous environment. This can be done with existing technology rather than waiting for telerobots to incorporate AI to perform "smart" autonomous operations. These two innovations (TSEE and 6-DOF tracking/orientation) are the key technologies to enable more sophisticated use of telerobotics, not just teleoperations, sooner in Space Station work, ground aerospace launch processing, and hazardous petrochemical or nuclear safing operations worldwide. This product can also enhance robotic applications in NASA and industry without redesigning existing valve panel facilities for robotics. It will provide safer and less serial-time operations in hazardous environments.

ORBITER RADIATOR DAMAGE INSPECTION

KSC (NASA) is working on a joint project with Lockheed (Kennedy and Palo Alto organizations) where Lockheed is developing a sensor system to examine and inspect the Orbiter radiators for damage (delamination and meteorite dings) while NASA is developing a robotic mechanism to transport the Lockheed sensor over the complex contours and 10.5 by 60 foot surface of the radiators. A prototype production system and future production models will be tested, certified, and installed in the Orbiter Processing Facilities (OPF) at KSC.

The objective of the project is to decrease the amount of time it takes to process an Orbiter before each mission. The efficiency of the radiators to dissipate radiant heat energy is dependent on their surfaces being clean and damage free. It presently takes 16 people 24 hours to inspect the Orbiter radiators to determine damage (dings, scratches, impacts) prior to continuing other work in the OPF. A robotic inspection system should reduce this to two people in 3 hours and provide accurate repeatable trend data and a quality inspection. Currently the operation is performed with an XYZ mechanism over the Orbiter that moves "buckets" large enough to carry men and equipment. The buckets are driven by highly skilled personnel. The visual inspection is performed using the naked eye and observations are recorded on a log sheet. This does not provide an accurate permanent record of damages and their locations. Small imperfections may be overlooked. This is a very uncomfortable, task-intensive, repetitive job.

The radiator panels are normally inspected twice during all OPF flows. The first inspection, postflight zonal inspection, is performed to detect any damage to the radiator panels that may have occurred in flight. The second inspection, immediately prior to payload bay closeout for flight, is performed to detect any damage that may have occurred during OPF processing. Other inspections are required if the radiators are removed from the payload bay doors and placed in storage.

Operations personnel are primarily concerned with anomalies that fall into two categories: (1) damage to the radiator panel surface silver Teflon tape and (2) damage to the radiator aluminum facesheet and honeycomb sandwich core. To date, inspection of the radiator panels has focused on only damage visible to the normal unaided human eye. Once a nonconformance condition is detected and documented, other techniques are performed to assess the damage severity.

Anomalies to the radiator surface include scuffs, scratches, tears, discoloration, delaminations, and bubbles in the silver Teflon tape. An automated inspection device must detect these anomalies and differentiate one from the other. The inspection tool will need to maintain a memory (damage log) of each radiator panel to determine if an anomaly was "new" or had been previously assessed and documented. Because most damage to the thermal control coating is not repaired, it is essential that the damage log provide a means for continual update and for referencing the cummulative surface area of minor unrepaired damage.

Anomalies to the radiator aluminum facesheet and honeycomb core include indentations (dents), scratches, pin holes, punctures, gouges, and meteorite strikes. The automated inspection device will be able to detect these anomalies and differentiate between defects that penetrate the facesheet from those that impact the facesheet but do not penetrate. As before, the inspection tool will maintain an automatically operated memory (damage log) of each radiator panel to determine trends and to quickly locate repair areas that are to be processed at a later date.

By automating this operation, most of the personnel will be relieved of their inspection duties so their talents could be used to perform other jobs more demanding of their skills. It would also provide a quickly accessible permanent record of radiator inspection data (damage assessment and precise location), reduce the amount of paperwork required to get the job done, minimize setup time to get ready for inspection, and provide expansion capabilities so that other functions could be performed on the radiators with the automated mechanism (e.g., cleaning). The overall goal is to achieve an efficient and less expensive operation.

Automatic Radiator Inspection Device (ARID)

The ARID mechanism was an evolutionary design culminating in a 4-degree-of-freedom robot (see figure 9). Lockheed originally envisioned that the ARID would be a camera traveling on a contoured beam shaped like the radiators, transported along two rails that lay along the edges of the Orbiter radiators (see figure 10). However, there were four major problems in implementing this configuration:

(1) Two of the major design goals were to not impact existing operations and to install the ARID without major modifications to the facility. The first simple concept would have led to producing a large mechanism (10 by 60 by 10 by 60 feet) which would have to be picked up by an overhead crane and moved into place with one transport rail latched in place to the Orbiter Processing Facility (OPF) platforms and the other transport rail suspended over the Orbiter's radiator hinges. This hanging structure would impact other operations that require the "buckets" to be moved. It would also have to be removed to open and close the radiator doors or to lower access platforms above the radiator to allow personnel access to payloads inside the Orbiter bay. Because of these facility impacts, the rail structure would have impacted serial flow, not reduced it

(2) A simple rod-like transport rail over the radiators was not feasible because of the 65-foot length required to span the cargo bay without resting on the Orbiter hinges. A truss assembly would have been required to reduce bowing and sagging of the structure, making it heavier and more difficult to move.

(3) Trend photographs of an old image must be aligned with a new image for comparison. If the alignment is off by more than 10 pixels or if the light angles are different, then the image will look different in the vision system processor's eyes. The bulky mechanism would have to be fitted with alignment offsets to allow for x, y, and z, and orientation differences for each vehicle and to allow for imprecise placement of the Orbiter in the OPF. This would have further complicated the design and introduced complex manual alignment procedures.

(4) The forward radiator panel (see figure 10) has a 5.71-degree of slope (covering an area 7 inches high by 70 inches long). This eliminates a fixed track design from being able to access all four side panels.

During the design, it became readily apparent that all these problems could be minimized through a flexible robotic mechanism that could: (1) provide a quick, software-programmed "frame shift operation" eliminating parking offset adjustments, (2) be cantilevered from under access platforms located adjacent to the radiator panels and moved out of the way when necessary so as to not impact facility operations, and (3) be reprogrammed or be expanded to support other future changes. This robotic flexibility will be advantageous for future update of the system to allow for cleaning of the radiators by the ARID rather than by men hanging over the sides of buckets. In the near future, as the use of the robot becomes more of a standard operation, as people get more accustomed to using automation, and after the ARID proves itself to be reliable, a modification will be made so that it can actually clean the radiators by "hand rubbing" the delicate surface.

Figure 11 depicts the first in-house mechanism concept (a PPRP robot). It consisted of a 65-foot prismatic rail to traverse the length of the four radiator panels, a prismatic cart to reach from the outer edge of the door to the Orbiter's hinge, and another prismatic rack-and-pinion rail with a rotary joint to move the inspection device over the contoured surface of the radiators. There were also problems with this configuration that caused it not to be implemented.

(1) Grease drippings from the Thompson rails and the rack-and-pinion gears used in such a design would pose problems in cleanliness. For example, radiators are covered after they are cleaned allowing upper platforms to be lowered (see figure 12). If they were not covered, dirt or tiny objects could fall on them as personnel walk on the platforms to access experiments in the payload bay.

(2) A PPRP device would have created problems with the management of lines and cables as the cart moved back and forth.

(3) The required inspection path would carry the vision system payload 6 inches above the radiator surface (see figure 12). However, this configuration required many closeup pictures and would have produced an excessive amount of data. The time to process this data would not have significantly reduced operational timelines.

To solve the first two problems, the design quickly evolved into a PRRR mechanism (see figure 13) in which the cables could be routed within an enclosed space (bending at the joints rather than being dragged over the width of the radiators). Seals can be installed on the joints to eliminate grease drippings.

In order to solve the third problem, Lockheed (the designers of the payload) requested that NASA KSC (the robot designers) investigate designing a mechanism that would allow taking photos from 6 feet away. This design would require that the upper access platforms be raised and another revolute joint be installed to lift the nose of the device (refer to point A of figure 13). Also, it would have made it difficult to design the robot to allow for later modification so it can reach the radiator surface for cleaning. However, a 6-foot inspection proved to be too far away to obtain reliable photos even with zoom lens. Also, since the designers did not want the device to hit and damage the radiators if it fell, it was determined that a 24-inch inspection distance would be the required sensing position above the surface of the radiators. The link lengths of the robot were then designed so the robot could transport the sensor over the work envelope (described in figure 13) without hitting the radiators at its extreme lower position. A mechanical stop keeps the robot within this work envelope. The link lengths were designed so the stops could be removed later and still be long enough to reach the surface of the radiators for any future tactile tasks.

During the development of the ARID by NASA, Lockheed was performing prototype sensor developmental tests. These tests revealed that a small vision sensor could be built (but it had to be aligned with the radiator panels to pick up all defects). Therefore, to use it on the 5.7-degree sloped surface of panel one, the smaller sensor required that the robot be designed with a fifth rotation axis (PRRRR configuration). At this time, the design of the robot was 90 percent complete, and this modification would have added too much weight and control complexity. A larger sensor could be built that did not require precise orientation, but the larger vision sensor would not fit into the operational envelope required to tuck the robot into a stowed position. The final solution was to build a "crook" or offset in the robot to enable use of the larger sensor (refer to figure 9).

The tolerances are so tight in the OPF that the robot arm must be moved in order to open and close the doors of the Orbiter. It was first thought that a hinge could be added in the area indicated by point B of figure 13. This, together with a revolute joint at point A, would allow the arm to be moved to a horizontal plane and then, using the hinge, be moved out of the way and stowed in parallel with the track itself. However, the payload bay doors are not strong enough to be opened on the ground in zero gravity conditions; therefore, when opening and closing the radiator doors, a strong-back device is attached to the outside surface to provide structural support. CAD drawings of the facility revealed that this strong-back device would pass at two places through the only area in which the prismatic track could be mounted (see figure 9). It would also pass through the robot arm if it were stowed in front of the track.

Two solutions were needed before the robot could be installed without modifying structures within the facility. First, an area (or "cubbyhole") at one end of the payload bay doors was located to store a folded-up arm (refer to the crosshatches in figure 13). If the arm was not foldable, it would have to be removed to open the doors. There was just enough space available in the cubbyhole to redesign the arm and joints to fit and still enable reach capability for a 24-inch nontactile inspection path and a 6-inch cleaning path. Second, in order to eliminate all facility impacts, the track was sectioned so that two short removable pieces could be quickly unbolted and slid back. These two modifications allowed the strong back passage through the track's operational envelope and resolved the final problems during the design evolution of the inspection robot.

Redundancy Requirements

Since this is the first robot that will be installed next to flight hardware (and especially since it hangs over the radiator doors), the reliability of the robot is extremely important. Recently, several suspended loads at various NASA facilities have had mechanical failures that caused them to drop onto a Spacecraft or flight hardware. If this occurs at the launch pad, not only could an $80 million Spacecraft be damaged, but it would cause an aborted launch. This produces expensive consequences resulting in large amounts of serial-time/money to replace and repair. The existing inspection "buckets" have even collided with the radiator doors and bent them upward. The robot has been designed to not hit any of the radiator panels even in worst-case parking conditions. However, the robot was also designed (for future use) where it could actually touch the entire surface of the radiators. When, and if, the stops and mechanical constraints are removed, the robot control system must be so reliable that there can be no inadvertant moves to cause even "dings" or "scratches" to the sensitive surface. To prevent this from occurring, the ARID has been designed so that it contains complete control system component redundancy and additional mechanical constraint redundancy.

Electromechanical design of each joint includes redundant drive shafts, bearings, harmonic drives, brakes, transmission chains, encoders, and motors (see figure 14). They are sized for the load, torques, and the space available in the stowage position. Electrically, the control system includes redundant motor servocontrollers, redundant indexers, and two separate control computers. Mechanically, redundancy exists by adding a special cable/pulley configuration (not shown) at joints two and three that prevents the arm from colliding with the radiators even if both redundant brakes or motor drive chains fail.

Two motors in parallel could possibly be expected to "fight" each other, but there is a 400-to-1 transmission gear ratio and a small torsional compliance in the harmonic drives and development in the laboratory has resulted in synchronous control of redundant motors. Also, drive components are balanced relative to the load. Therefore, any noticeable motor torque imbalance caused by component failure will be detected and used to shut down the system. The internal motor resolvers are compared against duplicate incremental encoders to provide redundant sensor feedback to enhance troubleshooting by the two computers. Once the fault is identified, the operator can shut down one side of the drive, release its brake, and use the remaining drive to fold the arm back into its tucked position and stow it in its "cubbyhole." In the case of a failure that requires repair of the robot, this will allow the doors to be closed without having to dismantle the robot arm and will enable the operational workflow to continue. A master computer and the slave computer each compare its own calculated kinetic positions, cross-check positions with each other, and check sensor feedback before allowing parallel-commanded moves to continue. The redundant system will be used in a "fail-safe" mode. If one computer fails, the other computer is switched in as a lone controller allowing the operation to continue to a safe conclusion.

Additional computer interfaces include digital/discrete control of brakes, emergency stop, digital sensors, limit switches, a manual control pendant, and a Lockheed-developed vision system computer data bank. The teach pendant allows manual override and speed control, and provides individual control of the position of all joints of the robot.

Simplication is the key to the design. Links 2, 3, and 4 of the robot will basically operate in a plane. The arm will be positioned to a location, the joints will be locked, and then the arm will travel lengthwise down a panel (using only the motors in link 1's prismatic axis). Next, the revolute joints will be unlocked and the arm will be lowered to another position that extends outward from the Orbiter (refer to figure 13). Then the arm will be moved lengthwise again. Photographs will be taken without stopping as the system moves at 4 inches per second. After an automatic scan, the operator may wish to look again at an anomoly or take a "still" photograph. He then can either program that position for an automatic move or manually drive the robot to a point that requires more resolution.

CONCLUSION

The first two innovations discussed in this paper (6-DOF Target Tracking System and TSEE) were designed and developed by contractors (Adaptive Automation Incorporated, and Automated Dynamics Corporation, respectively) working from specifications and with close guidance from NASA. These innovations can be integrated to perform tasks involving orientation and panel operations by using existing technology to perform "smart" autonomous operations. The third innovation (an ARID robot) was designed by NASA engineers. Contractor support is being used to fabricate and install the robot, to detail the teach pendant, and to develop the vision sensor. NASA engineers and co-op students performed the detail design of the joints, links, controls, and kinematics; and they performed kinematic, static, and dynamic analyses of this newly developed robotic mechanism as an "in-house" NASA developmental project. It is significant to note that the ARID robot involves the application of complete control hardware redundancy to enable performance of work above and near expensive critical Shuttle hardware (see figure 15). There are applications in which a robot would not normally be used because of "reliability" or in situations where an inadvertent move could result in damage to expensive components or even more expensive personnel. In such applications, the consumer marketplace may wish to explore commercialization of these component redundancy techniques. The three innovations discussed in this paper are applicable to Space Station work, ground aerospace launch processing, industrial safety, and hazardous petrochemical or nuclear safing operations worldwide.

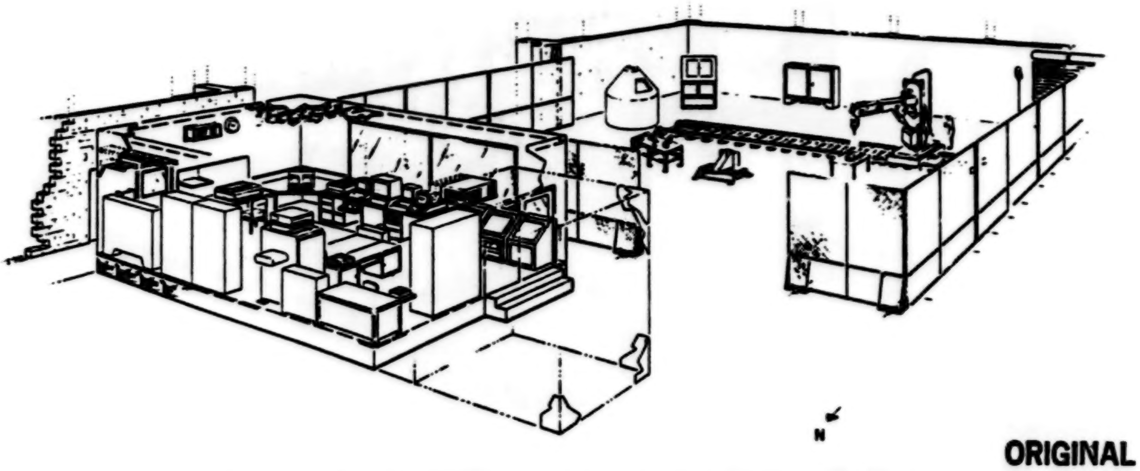


Figure 1. Robotics Applications Development Laboratory (RADL)

ORIGINAL PAGE IS
OF POOR QUALITY

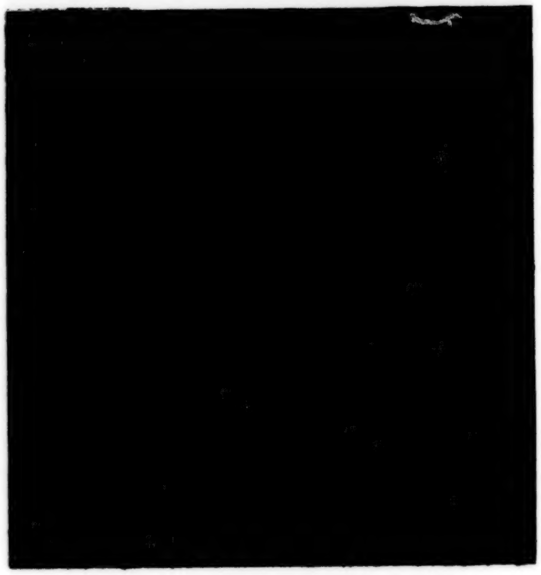


Figure 2. Remote Umbilical Mating
With Dynamic Simulator Target

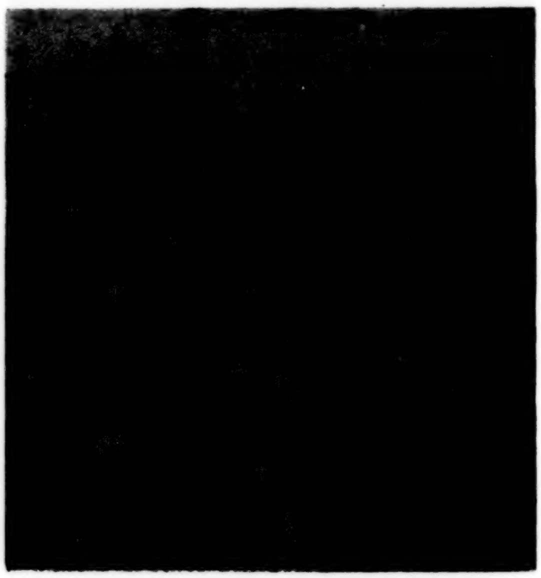
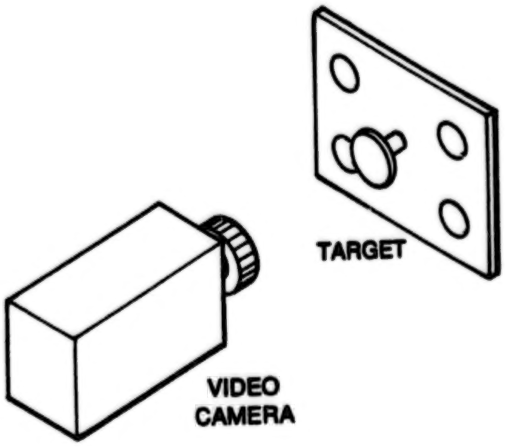


Figure 4. 6-DOF Target Discrimination
in the RADL



The Five Bright Circles of the target are positioned in such a way that the video images of them can be processed into data on the position and orientation of the target relative to the camera.

Figure 3. Tech Brief

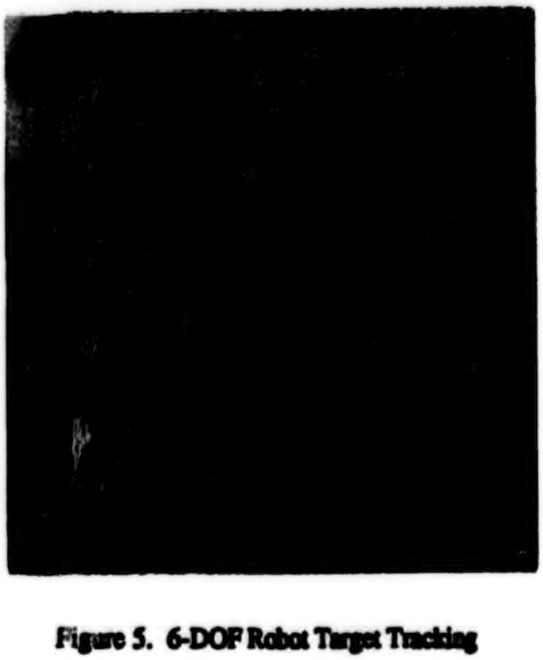
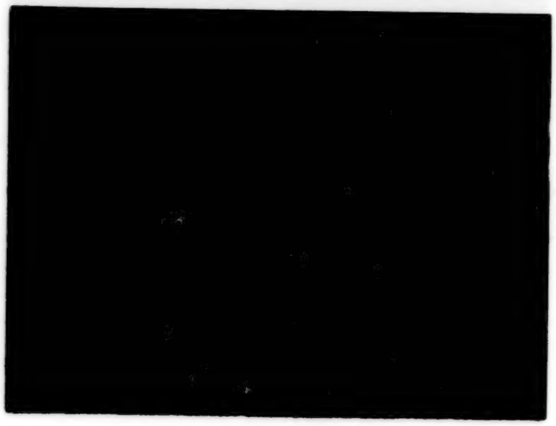


Figure 5. 6-DOF Robot Target Tracking

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH



ORIGINAL PAGE IS
OF POOR QUALITY

Figure 6. Small Phase I TSEE

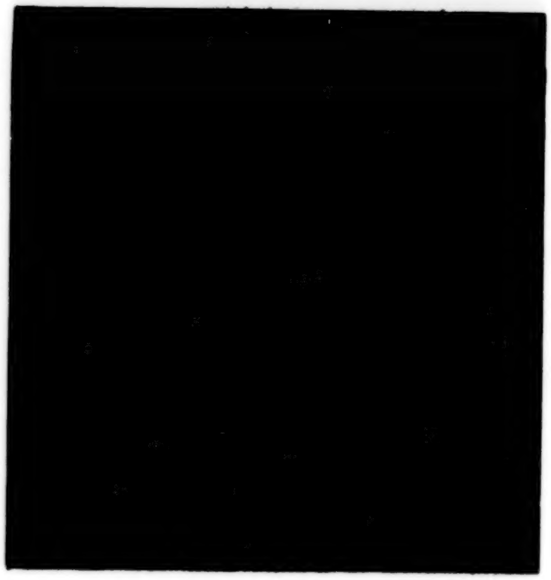


Figure 7. Large Phase II TSEE

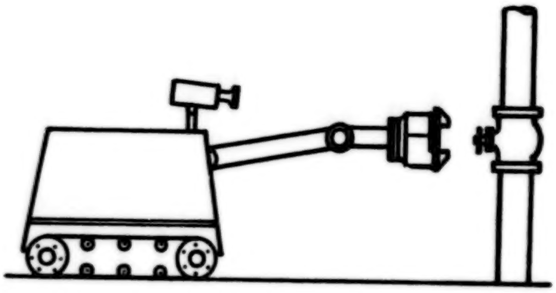


Figure 8. Remote Operation of TSEE

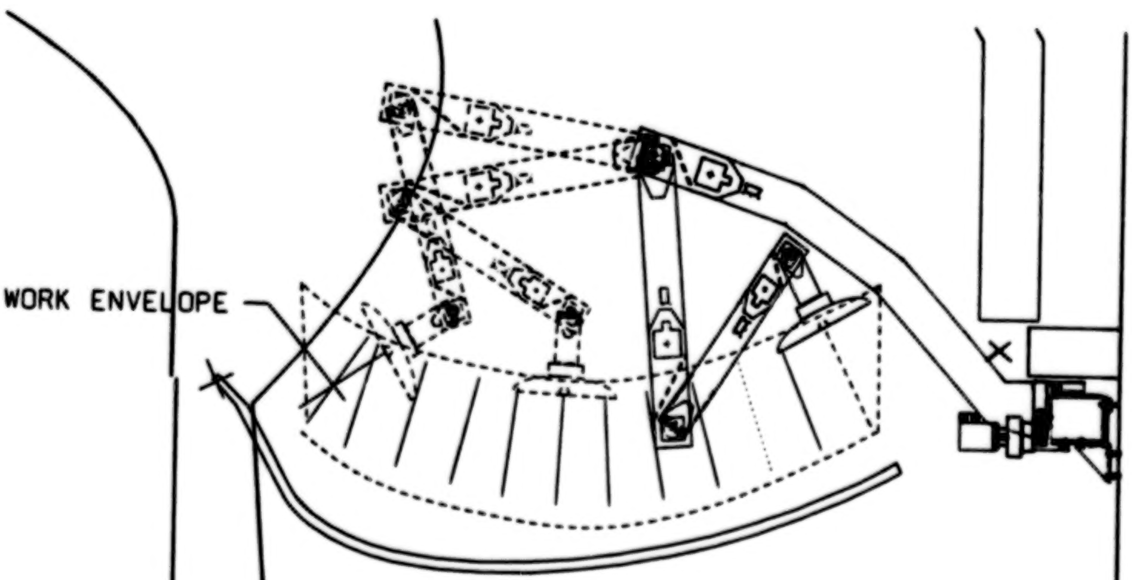


Figure 9. Radiator Inspection Robot

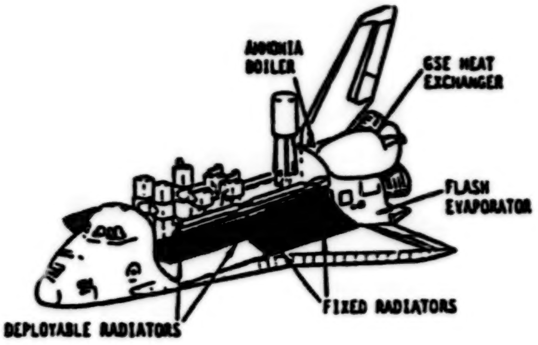

AMMONIA BOILER
GSE HEAT EXCHANGER
FLASH EVAPORATOR
FIXED RADIATORS
DEPLOYABLE RADIATORS

Figure 10. ARID Work Envelope

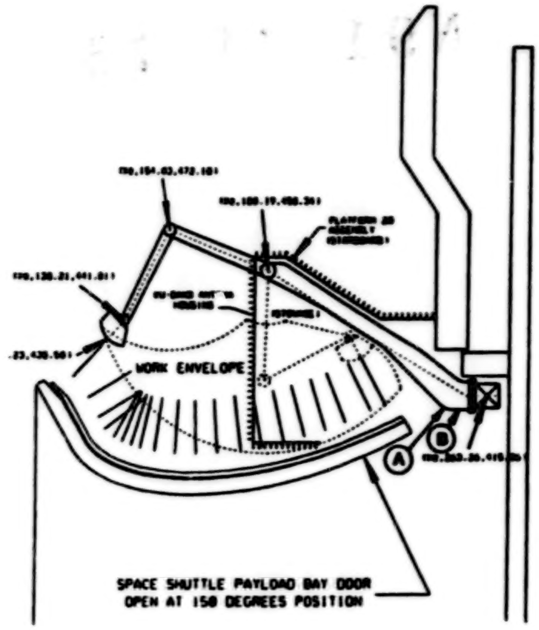

WORK ENVELOPE
A
B
SPACE SHUTTLE PAYLOAD BAY DOOR
OPEN AT 150 DEGREES POSITION

Figure 13. 24-Inch Inspection Work Space

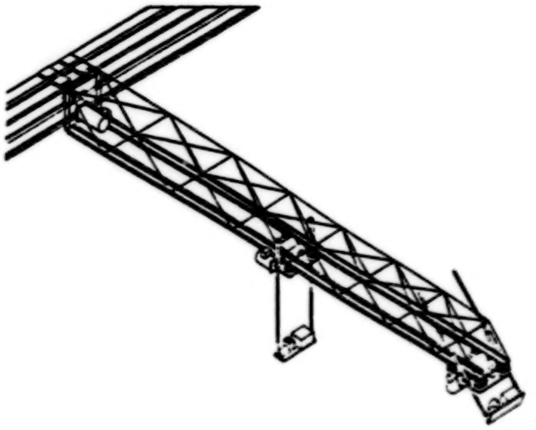


Figure 11. Early Concept

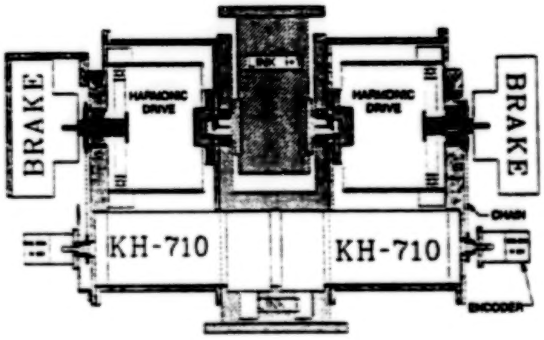

BRAKE
HARMONIC DRIVE
HARMONIC DRIVE
BRAKE
KH-710
KH-710
CHAIN
ENCODER

Figure 14. Redundant Drive Components

ORIGINAL PAGE IS
OF POOR QUALITY

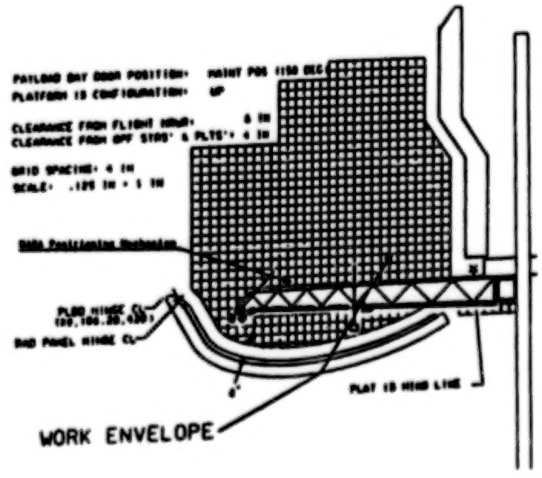

PAYLOAD BAY DOOR POSITION: MAINT POS (150 DEG)
PLATFORM IS CONFIGURATION: UP
CLEARANCE FROM FLIGHT HDWR: 6 IN
CLEARANCE FROM OFF STBD' & PLTS': 4 IN
GRID SPACING: 4 IN
SCALE: .125 IN = 1 IN
PLBD HINGE CL
RAD PANEL HINGE CL
0°
PLAT IS HING LINE
WORK ENVELOPE

Figure 12. Concept for 6-Inch Inspection

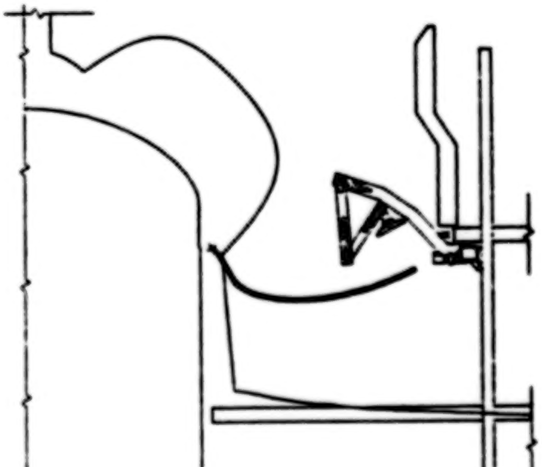


Figure 15. Final ARID Configuration

N91

24076

UNCLAS

N91-24076

THE TRANSFER OF TECHNOLOGYTO MEASURE SKIN BURN DEPTH IN HUMANS

William T. Yost and John H. Cantrell
NASA Langley Research Center
Hampton, VA 23665

INTRODUCTION

This technology transfer has two origins. The idea of using ultrasound to measure burn wound depth came from foundational work performed by Dr. Cantrell and others (1, 2) at Oak Ridge National Laboratory. The ultrasonic instrumentation technology came from an area which was developed at NASA Langley Research Center for improved resolution of ultrasonic nondestructive evaluation equipment to locate cracks in metal structures. The combination of the work in both areas has led to the development of a device that can detect burn-wound depth in humans.

While details will vary among technology utilization projects, this project points to a model paradigm for the case where the technology transfer involves measurements with subtle and, initially, often hidden meanings and interpretations. It is absolutely essential to under stand the meaning of the measurements taken in the area of the transfer, and this must become a key ingredient in the transfer process.

We begin with a brief, somewhat simplified description of skin tissue. Skin tissue is the external covering of the body, whose functions include protection against injuries and parasitic invasion, and the regulation of body temperature. It is made up of essentially two layers. The outermost layer, the epidermis, consists mostly of layers of cells which are flattened toward the outside. The innermost layer, the dermis, contains a variety of structures, such as nerves, nerve endings, sweat glands, blood vessels, and hair follicles. Capillaries provide nourishment and essential fluids for the skin. An important structural component for skin is collagen, a very large protein molecule that in the normal state is coiled. We start with the fact that skin is a medium through which ultrasonic waves can pass.

THE FUNDAMENTALSTUDY

Central to any application of ultrasonic technology is a property of the medium or propagation called acoustic impedance. This is defined as the product of mass density and the sound velocity as shown in the equation,

$$Z = \rho v \qquad (1)$$

If the acoustic impedance remains constant along the propagation path of the ultrasonic wave, the wave will continue to propagate through the medium. If, however, the acoustic impedance changes abruptly or discontinuously, the situation shown in Figure 1 occurs.

An incident wave is considered one that propagates from left tot right in medium 1 and is characterized by a given constant value of the acoustic impedance. The wave propagates through medium 1 and a second medium 2 which is characterized by a different value of the acoustic impedance. The abrupt change is the acoustic impedance at the interface causes part of the incident wave to be reflected back through medium 1 toward the ultrasonic source. The remaining part of the wave is transmitted through the interface into medium 2. There is interest only in that part of the wave reflect back through medium 1 to the ultrasonic source since in the assessment of burn depth there is access only to one side of the skin tissue.

An experimental arrangement as this, whereby one sends out an ultrasonic pulse and "listens" for reflections or echo signals from interfaces is descriptively called the ultrasonic pulse-echo technique. The amount of the incident wave intensity reflected in the echo signal is directly proportional to the square of the difference in the acoustic impedances,

$$I_r = \left(\frac{Z_1 - Z_2}{Z_1 + Z_2}\right)^2 I \quad , \qquad (2)$$

Where I_r is the intensity of the reflected wave, and I is the intensity of the incident wave, and the subcripts designate the medium. For the case of burned tissue, medium 1 represents necrotic tissue and medium 2 the underlying viable tissue. If the differences in the acoustic impedances of the necrotic and viable tissues are sufficiently large to obtain a signal from the necrotic/viable tissue interface that is strong compared to background reflections then the technique would provide a potentially quantitative measure of burn depth.

Another acoustic property of a medium that can affect the intensity of a received ultrasonic signal is the attenuation of the medium. Attenuation causes a decrease in intensity as the ultrasound travels through the medium. It depends on various factors, including ultrasonic frequency. The effect can be written as

$$I = I_0 e^{-2\alpha x} \quad , \qquad (3)$$

where α is the attenuation coefficient, x is the thickness traversed, and I_0 is the intensity of ultrasound at x = 0.

The Initial Tests

An ultrasonic pulse-echo system of sufficient resolution was assembled to see if, indeed, it would be possible to use ultrasound to investigate burn wound depth. Figure 2 (right side) is an ultrasonic reflection from a burn wound site on a Yorkshire pig. A transducer, sitting atop a column of gel used as a medium to transmit the ultrasound between the transducer and the skin, is used to interrogate the burn site. As can be seen from the ultrasonic trace, a strong echo is received from the interface of the ultrasonic gel and skin surface. After a series of closely spaced reflections from the necrotic tissue, a reflection from the burn interface is observed, as are reflections from the dermis-fat interface and other interfaces within the skin. For comparison the ultrasonic echo is juxtaposed with the histological cross-section made from excised tissue from the same wound.

One can see the effects of both reflection and attenuation in the ultrasonic trace. The reflection from the gel-skin surface is the first reflection received and the largest. Other reflections or echoes received decrease in prominence partly because as they traverse greater distances in the skin, they become weaker.

An analysis of the data collected compared as shown in Figure 2 gave excellent agreement between burn depth measured by histologic cross-section and measurements of the ultrasonic reflection from the necrotic-viable tissue interface. The average deviation between the two sets of measurements is 2.2% with a maximum deviation of less than 5%. To within the uncertainties of the measurements, the ultrasonic technique can be used as reliably as histologic sectioning in the measurement of skin-burn depth in porcine skin, which physiologically similar to human skin.

The Theoretical Model

Other information about the nature of a burn wound was also obtained from constructing a theoretical model of the burn process (3). Separate heat conduction equations were written for the necrotic region and the viable region separated from the necrotic region by a plane. The skin surface was assumed to be held at 100° C, while the initial temperature of the skin was 35° C. When experimental data was substituted into the model, the temperature of the interface between the necrotic region and the viable region was calculated to be 65.3°C. This temperature is in agreement with the asymptotic temperature measured by Moritz and Henriques (4, 5) for the time-temperature relationship for thermal necrosis of porcine skin.

Also calculated from this model was the energy of transition of the collagen. A value of 11.7 cal/gm of collagen in skin tissue was obtained. The value of 11.7 cal/gm is consistent with other measurements of transition energies (enthalpy changes) of collagen from other sources (6-9). This is still another indication that the collagen phase transition plays a key role in burn necrosis.

Further studies in the role of collagen in burn injury were conducted by others. One of significant importance to our work was done by Bartos (10), who used an optical polarization technique to measure effects on the collagen in the burn wounds in Wistar rats to characterize thermal injury. From his studies he showed that at the collagen shrinkage temperature (63-66° C) collagen fibers lose their birefringence. He further notes that "collagen fibers destroyed by heat can no longer fulfill their physiological function; the corium becomes necrotized, and has to be removed and replaced by new connective tissue."

The role of collagen in skin necrosis is a primary one: As the skin necrotizes, the collagen undergoes a phase transition, or "melts," resulting in a difference in skin density between the viable state and the necrotic state. This gives rise to a difference in acoustic impedances between the two states. Examination of Equation 2 shows that an ultrasonic reflection occurs at this interface, which is the basis of the acoustic reflection shown and marked in Figure 1. According to Bartos this interface also divides the necrotic skin, which cannot support the process of healing, from the viable skin where healing processes can occur. This explains why the ultrasonic reflection from the necrotic-viable skin interface coincides with the histologic results.

The Role of Other Reflections

Other significant reflections from interfaces in skin are found. Reflections from the dermis-fat interface and the interface between the ultrasonic couplant (gel) and the from surface of the skin can be used in the process to correctly identify the reflection from the necrotic-viable skin interface. Using reference data for the acoustic impedance, we calculated the relative intensity of the reflections from each of the interfaces. Also of importance to the design was a knowledge of the range of attenuation in skin. According to Equation 3, the signals attenuate or decrease as they pass through the skin. The strength of the attenuation is expressed as α : the larger α , the greater the effect. A literature search was conducted to determine the attenuation of skin. Since attenuation varies with frequency, we also researched the attenuation's frequency dependence for skin.

The results of these investigations and calculations lead to the following prediction: if one applies a variable gain ultrasonic system in just the right amount to compensate for the attenuation of the ultrasonic signal as it passes though the skin, then the three major reflections (gel-skin surface interface, necrotic-viable skin interface, and the dermis-fat interface) will appear as reflections of nearly equal height (A specialized circuit to accomplish this is called a time-gain compensation circuit, or abbreviated TGC.). Consequently, by adjusting the parameters of the TGC one could bring the signal level of the reflections from the gel-skin surface and the dermis-fat interface to the same height. If set up in this way, the appearance of the echo from the necrotic-viable skin interface would always appear to be approximately the same height as the other two reflections despite the burn thickness. This is a very important point, since normally reflection intensities decrease the deeper into the skin that the necrotic-viable skin interface is located. Other reflections can consequently confuse the interpretation, unless an easily reproducible criterion is established. The success of making a clinically useful instrument hinged on this point. Without a means of clear and unequivocal interpretation, the instrument would be essentially useless in a clinical setting.

THE INSTRUMENTS REQUIREMENTS

Our first consideration was that the frequency would be high enough that the resolution would be better than 100 microns in skin tissue. This determined the minimum frequency response needed from the transducer. Once determined, the decision of the frequency characteristics of the system electronics could be made. The estimated range of attenuations found in the literature for measurements in skin helped to determine the overall gain requirements.

The first consideration was whether a commercial medical ultrasound machine existed with the requisite frequency response and TGC that could be used. All that we tested indicated that no commercial instrument on the market at the time could meet the requirements without substantial modification. We decided to examine the instrumentation available for use with nondestructive evaluation of material. We were able to purchase a basic pulse-receiver unit with a 60 MHz bandwidth, which we modified with a time-gain compensator of in-house design. The frequency response of the assembled unit was better than 50 MHz, with accurate and adjustable, truly exponential time-gain compensation throughout the frequency range.

The output from the modified pulser-receiver was connected to a storage oscilloscope to complete the A-scan ultrasonic system. The display on the oscilloscope face was an important consideration. We wanted the scan to cover a sufficient portion of the screen in order that measurement accuracy would not be adversely affected. We developed a circuit that triggered the oscilloscope upon receiving the reflection from the gel-skin interface. A delay line was added to the system in order that the beginning portion of the reflection from the gel-skin surface interface could be observed. Therefore, when the system was appropriately set up, the pulse from the gel-skin interface aligned precisely at the left edge of the screen graticle. The other pulses appear approximately as tall as this pulse, as shown in Figure 3. When the system was properly adjusted, and a measurement was to be made, a special foot petal and circuit was designed to signal the oscilloscope, causing the trace to be stored. A permanent record was made with a Polaroid camera attached to the face of the oscilloscope.

THE INITIAL TESTING OF THE INSTRUMENT

The initial tests were conducted on skin burns on anesthetized pigs. A surgeon scald-burned a small region on the skin, made an ultrasonic measurement, and excised a portion of the burn region for histological study, under the direction of Dr. Robert Diegelmann at MCV. The results were encouraging, but the surgeon had considerable trouble in maintaining perpendicularity of the transducer to the skin surface, a necessary condition to obtaining good reflections from the burn site. A special cylindrical column made of lucite that could be filled with gel was built. When used with the transducers, it placed the surface of the skin at the optimum location and alignment to obtain the best traces. We also began to use slightly lower frequency transducers that were focused, thus making perpendicularity less critical.

After arrangements were made, we established a team of NASA personnel to take measurements on three patients who had been admitted to the burns unit at The Medical College of Virginia (MCV). The procedure was under the supervision of Dr. B. W. Haynes, director of the burns center at MCV, and his staff. Dr. Haynes provided diagnoses, against which we could compare our measurements. In every case, our measurements were in agreement with his diagnoses, both of which were subsequently confirmed by the healing process. We felt that the concept was proven to be useful in the clinical setting, and the instrument and concept were ready to extend to the manufacturer.

THE TRANSFER OF THE TECHNOLOGY

The test site again provided valuable assistance during this phase of the project. For example, Dr. A. Marmarou of MCV critiqued the prototype that we used, and made a number of valuable suggestions. One suggestion included the conversion of the system to a scanning device, enabling the surgeon to obtain an ultrasonic cross-sectional view (B-scan) of significant regions of a burn site.

With help from Mr. John Samos and Dr. Franklin Farmer of our Technology Utilization Office, and in cooperation with Dr. Harold W. Adams from Langley's Office of Chief Counsel, Westminster Medical International Corporation was identified as the firm to refine, build, and market an instrument for burns depth measurement. During the time interval of this study, they had developed a high frequency B-scan ultrasonic instrument, which they interfaced to a personal computer for display and data analysis. Working closely with MCV staff and NASA Langley Research Center, the company has made further refinements in the transducer scanner design, data analysis algorithms, etc., in an effort to develop an market a user-friendly instrument for the measurement of burn wound depth. Details of the instrument are available from the company, whose address is 9 Ramland Road, Orangeburg, NY 10962. Their telephone number is (914) 365-2854.

REFERENCES

1. Goans, R. E., Cantrell, J. H., Jr., and Meyers, F. B.: Ultrasonic pulse-echo determination of thermal injury in deep dermal burns. Med. Phys. 4: 259-263, 1977.

2. Cantrell, J. H., Jr., Groans, R. E., and Roswell, R. L.: Acoustic impedance variations of burn-nonburn interfaces in porcine skin. J. Acoust. Soc. Am. 64: 731-735, 1978.

3. Cantrell, J. H., Jr.: Ultrasonic determination of thermodynamic threshold parameters for irreversible cutaneous burns. J. Acoust. Soc. Am. 73: 337-339, 1982.

4. Henriques, F. C., Jr., and Moritz A. R.: Studies of thermal injury: I. The conduction of heat to and through skin and the temperatures attained therein. A theoretical and an experimental investigation. Am. J. Pathol., 23: 531-549, 1947.

5. Moritz, A. R., and Henriques, F. C., Jr.: Studies of thermal injury: II. The relative importance of time and surface temperature in the causation of cutaneous burns. Am. J. Pathol., 29: 695-720, 1947.

6. Finch, A., and Ledward, D. A.: Shrinkage of collagen fibers: A differential scanning calorimetric study. Biochem. Biophys. Acta, 278: 433-439, 1972.

7. Haly, A. R., and Snaith, J. W.: Biopolymers, 10: 1681-1699, 1971.

8. Lim, J. J.: Transition temperature and enthalpy change dependence on stability and destabilizing ions in the helix-coil transition in native tendon collagen. Biopolymers, 15: 2371-2383, 1976.

9. Kuntzel, A., and Doehner, K.: Agnew. Chem., 52: 175, 1939.

10. Bartos, F.: Changes in skin collagen polarization - Optical properties after contact burn. Acta Chirurgiae Plasticae, 21: 50-55, 1979.

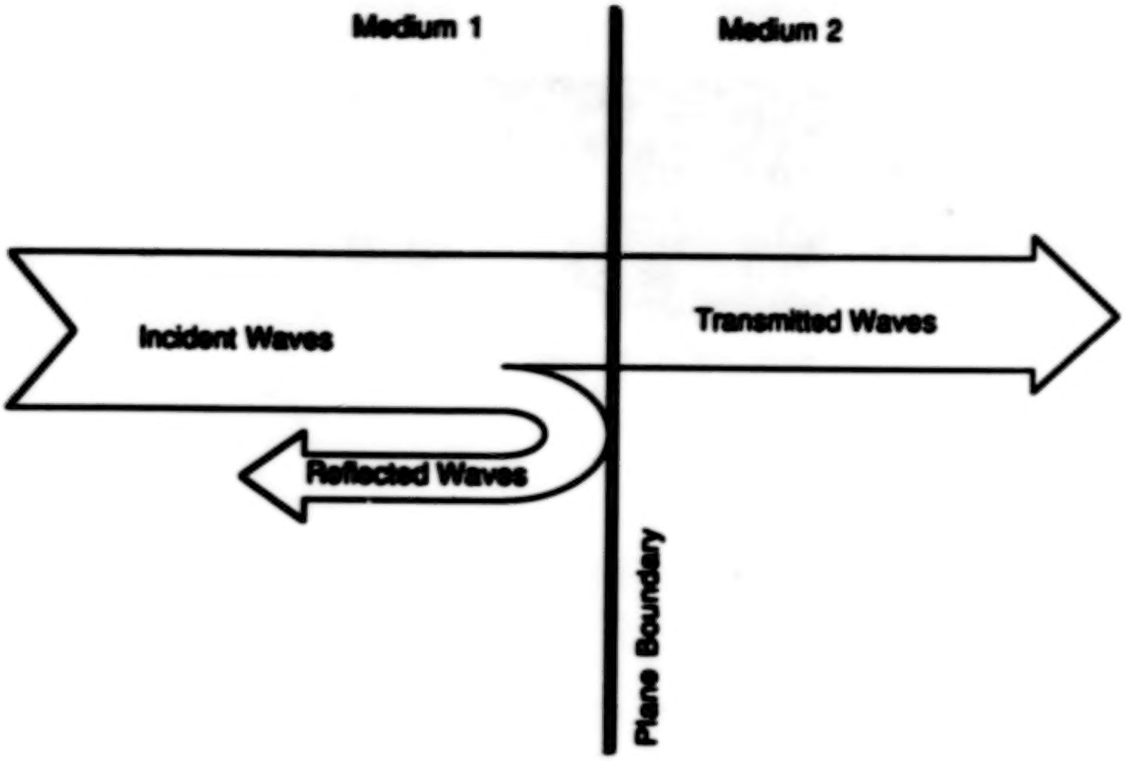

Medium 1
Medium 2
Incident Waves
Transmitted Waves
Reflected Waves
Plane Boundary

Figure 1. Reflection and transmission of sound waves incident at a boundary between media of different acoustic impedances

ORIGINAL PAGE IS
OF POOR QUALITY

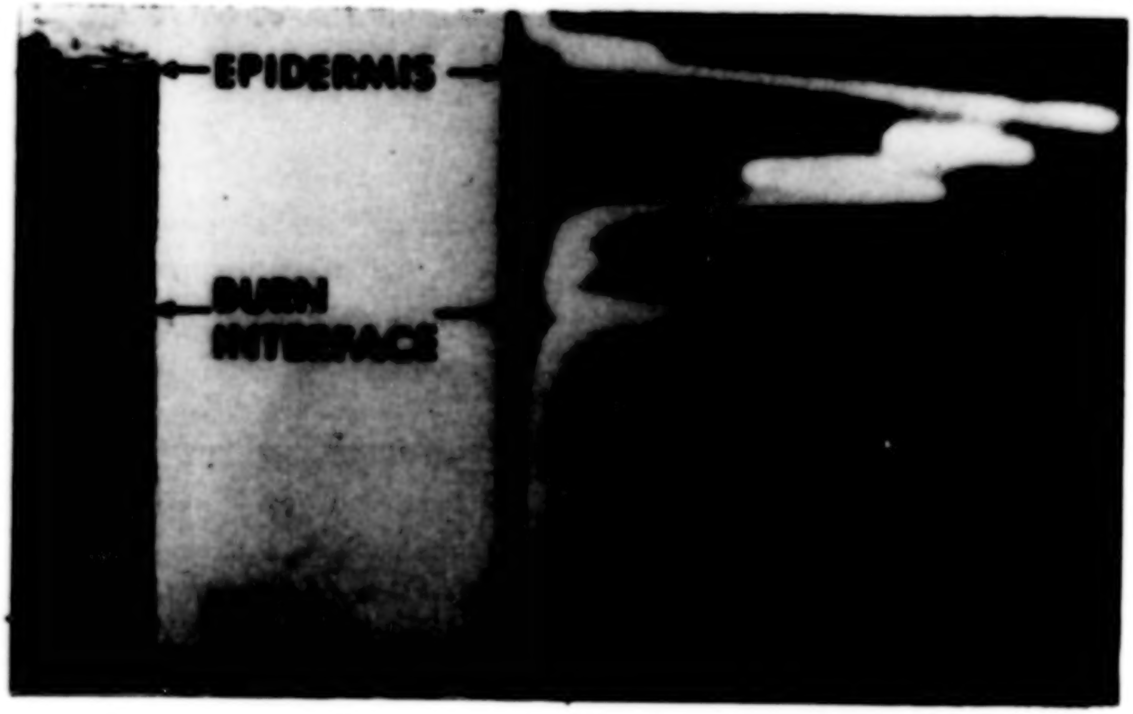

EPIDERMIS
BURN
INTERFACE

Figure 2. Comparison of a histological section taken immediately postburn with the corresponding ultrasonic pulse-echo data (A-scan trace)

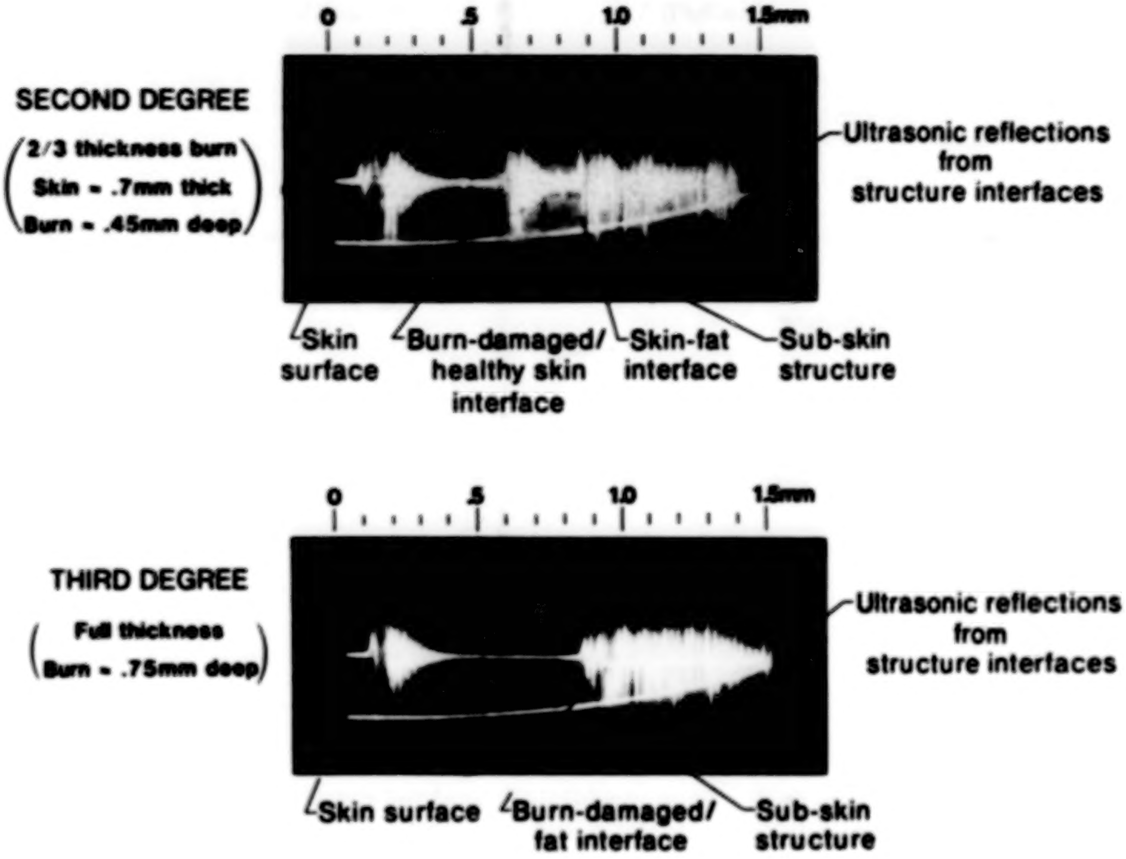


Figure 3. Display ultrasonic pulse-echo data (top trace) of second degree burn injury (dermal, deep dermal) and third degree burn injury (full thickness) in humans, with echos labeled. The bottom trace is the TGC control voltage.

ORIGINAL PAGE IS
OF POOR QUALITY

N91
24077
UNCLAS

N91-24077

Frequency Domain Laser Velocimeter Signal Processor

by

James F. Meyers
NASA - Langley Research Center
Hampton, Virginia 23665

and

R. Jay Murphy
Macrodyne, Inc.
Clifton Park, New York 12065

Abstract

A new scheme for processing signals from laser velocimeter systems is described. The technique utilizes the capabilities of advanced digital electronics to yield a signal processor operating in the frequency domain maximizing the information obtainable from each signal burst. This allows a sophisticated approach to signal detection and processing with a more accurate measurement of the chirp frequency resulting in an eight-fold increase in measurable signals over present high-speed burst counter technology. Further, the required signal-to-noise ratio is reduced by a factor of 32 allowing measurements within boundary layers of wind tunnel models. Measurement accuracy is also increased up to a factor of five.

Introduction

Laser velocimetry is a technique that nonintrusively measures the velocity of micron sized particles embedded in fluids or microscopic flaws in surfaces. The optical system creates an interference fringe pattern using two crossed laser beams, figure 1. The measurement volume created by the crossing beams is typically 0.2 mm in diameter and 2.0 mm in length resulting in basically a point measurement. As a particle passes through the fringes, it will scatter light from the lighted bands. A portion of this oscillating scattered light is collected by a photomultiplier tube where it is converted to an electrical signal. This signal, a burst of about a microsecond in length illustrated in figure 2, contains a chirp whose frequency is directly related to the rate at which the particle passed through the lighted bands. The burst can range from photon resolved to a complete analog signal up to 400 mV in amplitude. The signal can be masked by extraneous noise from collected photons originating from background light, laser light scattered from very small particles—too small to be individually realizable, and/or from laser light scattered from surfaces.

Originally laser velocimeter signals were processed using standard spectrum analyzers or specially built wide-band frequency trackers.[1] These frequency domain approaches worked well with continuous signals obtained from liquid flows, but the analog circuitry could not reliably process the individual signal bursts obtained from air flow measurements.[2] If a single signal burst could be detected, a timer can be used to measure the time between the first cycle of the chirp and the nth cycle. This time domain approach became the basis of the high-speed burst counter, now the standard signal processor for laser velocimeter applications.[3,4,5] While the high-speed burst counter provides good measurement capability, it suffers from two limitations. It uses only two points to determine the cycle time, and it requires the chirp to fit the classic model of a signal burst. The high-speed burst counter provides measurement accuracies in the 0.5- percent range with residual turbulence intensities (indicated turbulence intensity with no turbulence in the flow) in the neighborhood of 0.5- percent. Theoretical investigations show that turbulence intensity measurements below 1.0- percent are unreliable. The lack of low turbulence measurement capability and the requirement for an experienced operator to keep the measurement inaccuracies within the above ranges led to an investigation to develop new technology to reduce these limitations.

This investigation resulted in the following conclusions: (1) Increased accuracy was directly dependent on the number of measurement samples obtained during the duration of the signal burst. (2) Multiple thresholds or multiple bit measurements from an analog-to-digital converter (ADC) coupled with proper signal processing would remove the requirement for classic signal bursts. (3) Processing in the frequency domain provided increased efficiency with better measurement accuracy given the increased number of multiple bit measurement samples. Fortunately digital signal processing hardware is now available which eliminates the problems found with the previous analog frequency domain processing techniques.

A signal processor utilizing a high-speed 2-bit transient recorder for signal capture and a bank of adaptive digital filters with energy and/or zero crossing detection was developed via computer model.[6] This model was tested using real and simulated laser velocimeter signal bursts at various amplitudes under various signal-to-noise conditions. The results of these simulations indicated the device would increase measurement accuracy over the high-speed burst counter by five times from signals with as few as 150 photons per burst. The residual turbulence intensity was reduced by a factor of 2.5 and turbulence intensity measurements down to 0.2- percent were reliable. Minimum acceptable signal-to-noise ratio was reduced by a factor of 32 as compared to the high-speed burst counter.

Although the simulated signal processor yielded impressive results, the device was impractical to build. The overhead required to implement the digital filter bank would limit the processor to a maximum measurement rate of less than 10 measurements per second. Utilizing optimized digital signal processing large scale integrated circuits to perform fast Fourier transforms and changing the filter bank to analog circuitry, a practical frequency domain signal processor was constructed—the FDP-3100.

Desired System Characteristics

The design studies leading toward the computer simulation of the frequency domain processor, FDP, began by evaluating the characteristics of laser velocimeter signal bursts and by developing the desired system characteristics. The evaluation indicates that the signal bursts are of short duration and obey Poisson occurrence statistics with very low average rates. The signal amplitudes have little variation if the particles are all approximately the same size. The turbulence intensity parameter f'/F must not be allowed to exceed 20 percent to prevent frequency aliasing at 0 Hz.* The signal processor should contain a transient recorder to capture the signal burst with a triggering circuit designed to maximize the data acceptance rate while minimizing the number of false triggers. The transient recorder should contain a multi-bit analog-to-digital converter. The processing of the captured signal should be as rapid as possible to minimize processor dead time. Signal validation circuitry should be included to give confidence that the acquired and processed data are indeed from valid signals.

Studies indicate that a signal processor is usually waiting for a signal burst to arrive. If the presence of the signal burst can be detected and the waveform captured and stored in a digital memory, ample time would be available for processing prior to the arrival of the next signal burst. This forms the basic strategy of the ideal signal processor: A high-speed burst capture circuit followed by signal analysis circuitry furnishing data to digital feedback loops for gain and average frequency adjustments and for the measurement of the chirp frequency in each signal burst.

Although the signal bursts are digitized by other laser velocimeter signal processing techniques, the only true multibit digitization of signal bursts has occurred by using transient recorders with computer processing of the captured bursts.[7] The major advantage of the transient recorder is the ability to capture the entire burst without the assumption of symmetry or high signal visibility as in the other techniques. A secondary advantage of the transient recorder is its maintenance of amplitude information which could be input to circuitry for controlling signal amplification. The disadvantages of a transient recorder, however, are manual setting of signal amplification and sampling rate, and the slow transfer of the data to the external computer. If these problems can be overcome, this approach would seem to have a distinct advantage over classical laser velocimetry signal processing.

---

* When u'/U is greater than 20 percent, a Bragg cell is used to bias the signal frequency away from 0 Hz; thus f'/F is reduced below 20 percent.

Computer Simulation of the Proposed Frequency Domain Processor

A computer simulation of a frequency domain processor was developed based on the above characteristics.[6] This system utilized a high-speed transient recorder to capture the signal burst and a bank of seven adaptive digital bandpass filters to develop an energy histogram which was then interrogated to determine the chirp frequency contained within the signal burst. Digital feedback loops were included to control the gain circuits in the input amplifiers and to control the reference clock frequency to maintain an average digitizing rate of 10 samples per cycle.

The transient recorder was designed to operate in the manner of a shift register with the digitized signal continuously passing through the recorder, allowing the contents to be integrated to detect a signal burst. When a signal burst has been found, the shifting process would be halted and the data transferred to data latches to provide the input to the signal processor section. This technique allows processing the laser velocimeter output only when signal bursts are present. By using the amplitude and frequency information obtained from the signal burst, the amplification and sampling rate can be adjusted to maximize measurement accuracy. In practice, the system should not adjust the sampling rate while acquiring data, since this would require the output of the rate along with the desired measurement information. Thus the system should be designed to operate in two phases, a setup phase to establish the sampling rate and a data acquisition phase with the rate held constant. The information on signal amplitude can be used to continuously control an automatic gain circuit to keep the signal amplitude within the optimum bounds. This basic approach is illustrated in the block diagram in figure 3.

Although it would appear that the most straight-forward approach to determine the oscillation frequency of the captured signal burst would be to perform a fast Fourier transform, FFT, and to determine the location of the peak, this approach is not necessarily the most efficient for laser velocimeter signal burst processing. (As will be shown later, this thought is incorrect considering the new state-of-the-art digital signal processing large scale integrated circuits that have been optimized for FFT calculations.) The FFT performs many unnecessary calculations to determine the energy distributions for frequencies above and below the oscillation frequency, since it calculates the entire frequency spectrum. It would be much more efficient if only those frequencies surrounding the oscillation frequency were investigated. This ideal can be realized by using a digital bandpass filter bank with control of the sampling frequency during the above-mentioned setup phase to approximate the desired Fourier energy calculations. By using f'/F = 20 percent and ±3 standard deviations, 99.73 percent of the energy contained in the Gaussian frequency (velocity) distribution will be obtained for the maximum measurement conditions. Therefore, a bank of seven digital bandpass filters, each with a bandwidth of at least $f'/F = 0.2$ and a center frequency separation of 20- percent of the sampling frequency can be used to obtain the equivalent energy distribution as an FFT with an increase in computation speed, figure 4.

The use of digital bandpass filters offers several signal processing possibilities. The peak of the energy distribution passing through the filter bank can be determined by using curve fit or statistical procedures to yield the oscillation frequency. The filter passing the greatest energy for a given signal burst could be used as a narrow bandpass filter for a high-speed burst counter. The signal-to-noise ratio may be increased for low f'/F by changing the coefficients of the filters to reduce both the bandwidth and the frequency spacing. These possibilities may even be combined by using a control algorithm designed to maximize measurement accuracy of the input signals.

The resulting simulation was composed of a transient recorder with unfiltered automatic gain amplification, analog-to-digital conversion, shift register storage, and burst detection circuitry; and a computation signal processor including a bank of seven adaptive digital bandpass filters, and energy detection, zero crossing detection, and automatic frequency control algorithms. The processor would use the techniques of automatic frequency control and automatic gain control to keep the input signal bursts at the optimum sampling and amplitude points. The unit would be insensitive to input signal-to-noise ratio while processing individual signal bursts. The overall system function diagram is shown in figure 3 and with increased detail in figure 4.

Testing of the Simulated Frequency Domain Processor

The operation of the control algorithms and the characteristics of the frequency domain laser velocimeter signal processor were tested by means of computer simulation of laser velocimeter signal bursts and by modeling of the signal processor. The simulation of the signal burst was by means of Poisson shot noise models generated in the manner described in reference 8. These models could be adjusted form the photon-resolved regime, through photon pileup, to photomultiplier saturation. The model of the signal processor includes the analog-to-digital converter, the automatic gain control and the automatic frequency control circuitry; signal detection circuitry; digital filter banks; zero crossing detection circuitry; energy detection circuitry; and signal processing algorithms to be implemented in the controlling microprocessor firmware. A typical test consisted of generating a series of signal bursts with a selected average photon count, average signal visibility, average oscillation frequency, and standard deviation of oscillation frequency. The oscillation frequency for each signal burst was statistically selected based on Monte Carlo methods with a Gaussian probability distribution of signal frequencies based on the requested average and standard deviation. The theoretical signal burst is generated based on the selected oscillation frequency and input visibility function and used as the driver for generation of the signal burst based on the algorithm from reference 8, figure 2. This burst is then input to the model of the frequency domain processor for setup and/or processing. Once the setup phase is completed, 100 signal bursts are input to and processed by the frequency domain processor. The statistics of the measured results are calculated and compared with the statistics of the input ensemble used to generate the signal bursts. This direct comparison of data removes the assumption that the characteristics of the 100 signal bursts agree with the requested average frequency and standard deviation. A model of a high-speed burst counter was developed to serve as a performance comparison. It used 4-pole Butterworth low- and high-pass filters, double threshold signal detection, zero crossing measurement by using a 500 MHz reference clock, and 5:8 count comparison error detection. Again the statistics of the input ensemble of signal bursts were compared to the output statistics of the measurement ensemble from the high-speed burst counter. It is noted that signal bursts that did not satisfy either the double threshold detection check and/or 5:8 error detection were not included in the measurement ensemble, whereas all measurements were included in the results from the frequency domain processor.

The first series of comparison tests determined the measurement accuracy of f'/F for high-level signals (approximately 1500 photons per burst at a signal frequency of 25 MHz—example given in figure 2(b)) for 0 to 5 percent. The comparison of measured f'/F with the input f'/F as a function of mean frequency for the high-speed burst counter is shown in figure 5(a). The companion comparison for the frequency domain processor is shown in figure 5(b).

The effect of signal-to-noise ratio on the measurements were tested by reducing the average photon count per signal burst from approximately 3000 per burst to approximately 150 per burst at a mean frequency of 25.0 MHz. As shown in figure 6, the high-speed burst counter has a residual f'/F approximately 2.5 times larger than the FDP down to 300 photons per burst. As the input f'/F is increased from 0 to 1.0 percent (figures 7 through 12), the results from the FDP match the input at an f'/F of 0.2 percent with bursts containing at least 900 photons. Signals with photon counts down to 500 match at 0.3 percent and signals down to 300 photons match at 0.4 percent. The high-speed burst counter results did not match until f'/F reached 1.0 percent and then had a great deal of scatter.

Translation to the Real World

The translation of an idea to hardware has a different set of constraints than translation to computer simulation. Simulation allows a full range of processing and control algorithms where time is unimportant. However, the hardware version must be designed for simplicity and speed to be usable. Unfortunately the simulated frequency domain processor is neither. The digital filter bank is complicated with excessive overhead reducing the anticipated data rate to unacceptably low levels. The control circuits for automatic gain and triggering for the transient recorder require up/down counters running at 1.0 GHz which would be difficult to construct even using GaAs technology. The automatic frequency control circuit is not necessary when fast Fourier transform approaches are used. Although these casualties may seem to completely change the character of the processor, many of the elements in these approaches are still incorporated, but in different ways. For example, the digital filter bank has been converted to a bank of analog filters and placed in the transient recorder to serve as an integral part of the triggering circuit.

To help keep the hardware as simple as possible, which in turn keeps the costs down, the automatic capability of the simulated processor was removed and manual adjustments provided for the gain and threshold settings. The user is aided in setting these controls by a graphic display on the front panel which shows the captured signal burst (updated once per second with the latest captured burst), the triggering threshold level and the Fourier transform of the burst.

Other modifications include increasing the analog-to-digital converter from 2-bits to 8-bits for increased resolution and allowing the record length to be user adjustable from 32 to 4096 samples per burst. The increased resolution of the analog-to-digital converter is necessary to minimize harmonics in the frequency domain. The adjustable record length allows the user to choose from a high data acquisition rate, but at reduced measurement accuracy, using the minimum record length or maximum precision with 4096 samples per record if data rate is not important. The optimum compromise between acquisition rate and measurement accuracy was found through simulation to be 256 sample records.

The FDP-3100

The resulting hardware system, embodied as the FDP-3100 shown in figure 13, consists of the transient recorder and the signal processing section. The transient recorder contains an anti-alias filter, a variable gain amplifier, the analog-to-digital converter, the triggering circuitry, and the high-speed memory used as the shift register. The signal processing section consists of the digital signal processing chip with its program and data memories, and the input/output circuitry. The operation of the FDP-3100 can be best described by following a signal burst as it is processed.

The signal obtained from the photomultiplier in the laser velocimeter is conditioned by an anti-aliasing filter which removes high frequency noise attributable to the random arrival of photons (shot noise) and any harmonics from the chirp frequency. The conditioned signal is amplified by the variable gain amplifier whose gain is set from the front panel or controlling computer. The amplified signal is simultaneously passed to an analog filter bank for signal detection and to the 8-bit analog-to-digital converter, ADC, for digitizing. The output from the ADC is passed to high-speed static RAM operating as a shift register. The shift register can be set to capture from 32 to 4096 samples in a binary progression at the master clock rate. The analog filter bank divides the input bandwidth of 20 MHz into 16 sections with 4-pole bandpass filters. The bandwidth of each filter is 1/6th of the center frequency of that filter. The center frequency of each filter is set so adjacent filters overlap at their respective 3 dB points. These filters provide a signal-to-noise improvement of 5-6 dB to any signal burst passing through the bank. The output from each filter is compared to the selected threshold setting. If any output exceeds the threshold, a trigger pulse is sent to the shift register to latch its contents, thus capturing the signal burst.

The latched signal burst is passed to the digital signal processor, DSP, for conversion to the frequency domain using a fast Fourier transform. The resulting frequency spectrum is interrogated to determine if a signal burst was indeed captured. The user can select whether the peak ratio or the energy ratio validation algorithm is used. The peak ratio algorithm compares the ratio of amplitudes from the highest and next highest peak in the frequency spectrum to the user selected validation ratio. The energy ratio algorithm compares the ratio of energy contained in the highest peak and the total remaining energy to the user selected validation ratio. If the validation is successful, the measurement frequency is determined by a weighted statistical analysis of the peak in the spectrum. The result is then passed to the data acquisition system for analysis and storage.

Concluding Remarks

The development of a frequency domain signal processor for laser velocimeter applications has been described. The historical discussion showed that the early work was on the right track with frequency domain processing, but technology at the time did not provide the tools necessary to accomplish the task. The development of large scale integrated circuits and high-speed digital circuitry now makes a frequency domain processor possible. The development and verification of the approach was performed using computer simulation. While direct conversion to hardware was impractical in many cases, the basic approaches were. The resulting hardware system, the FDP-3100, increases the performance envelope of laser velocimeter signal processing while making it easier to obtain good quality measurements than present technology.

References

1. Huffaker, Robert M.; Fuller, Charles E.; and Lawrence, T. R.: Application of Laser Doppler Velocity Instrumentation to the Measurement of Jet Turbulence. [Preprint] 690266, Soc. Automotive Engineers, Jan. 1969.

2. Fridman, J. D.; Kinnard, K. F.; and Meister, K.: Laser Doppler Instrumentation for the Measurement of Turbulence in Gas and Fluid Flows. *Proceedings of the Technical Program - Electro-Optical Systems Design Conference*, Karl A. Kopetzky, ed., Industrial and Scientific Conference Management, Inc., c.1970, pp. 128-146.

3. Iten, Paul D.; and Mastner, Jiri: Laser Doppler Velocimeter Offering High Spatial and Temporal Resolution. *Flow - Its Measurement and Control in Science and Industry*, Volume One, Part Two, R.E. Wendt, Jr., ed., Instrument Soc. of America, 1974, pp. 1007-1013.

4. Lenert, A. E.; Smith, F. H., Jr.; and Kalb, H. T.: Application of Dual Scatter, Laser Doppler Velocimeters for Wind Tunnel Measurements. *International Congress on Instrumentation in Aerospace Simulation Facilities - ICIASF '71 Record*, IEEE Publ. 71-C-33 AES, Inst. of Electrical and Electronic Engineers, Inc., c.1971, pp. 157-167.

5. Asher, J. A.: LDV Systems Development and Testing. *The Use of the Laser Doppler Velocimeter for Flow Measurements*, W. H. Stevenson and H. D. Thompson, eds., Contract N00014-67-A0226-005, Purdue Univ., Nov. 1972, pp. 338-347. (Available from DTIC as AD 753 243.)

6. Meyers, J. F.; and Clemmons, J. I., Jr.: Frequency Domain Laser Velocimeter Signal Processor - A New Signal Processing Scheme. NASA TP-2735, Sep. 1987.

7. Griskey, R. G.; Balmer, R. T.; and Brinkley, J. L.: An Automatic Scanning LDV System. *Minnesota Symposium on Laer Anemometry - Proceedings*, E. R. G. Eckert, ed., Univ. of Minnesota, Jan. 1976, pp. 217-230.

8. Mayo, W. T., Jr.: Modeling Laser Velocimeter Signals as Triply Stochastic Poisson Processes. *Minnesota Symposium on Laser Anemometry - Proceedings*, E. R. G. Eckert, ed., Univ. of Minnesota, Jan. 1976, pp. 455-484.

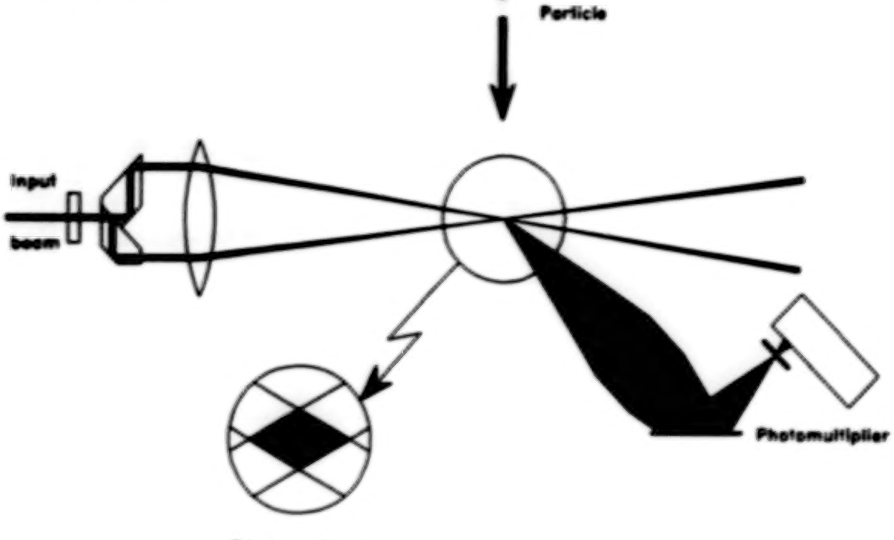
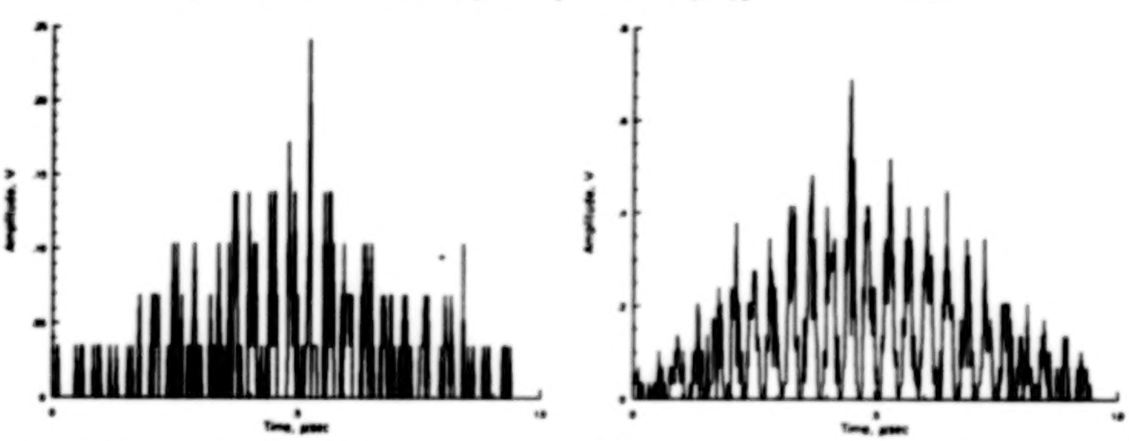

Particle
Input
beam
Photomultiplier
Fringe pattern

Figure 1. Schematic of a single component, fringe-type laser velocimeter.



(a) Signal burst containing 310 photons. (b) Signal burst containing 1500 photons.

Figure 2. Signal burst from fringe-type laser velocimeter with modulation frequency of 25.0 MHz.

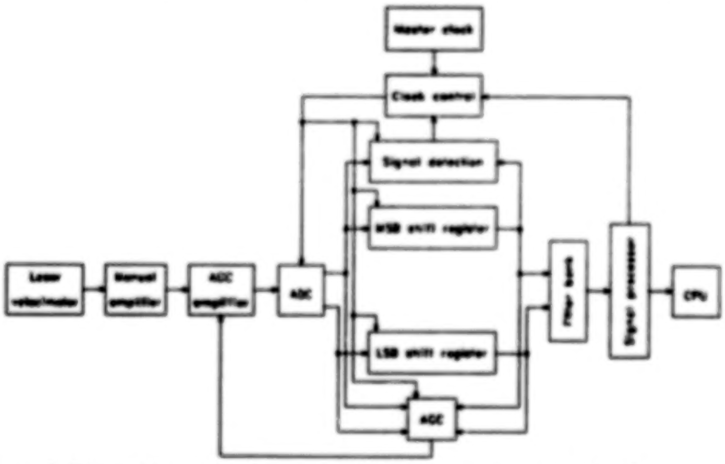

Master clock
Clock control
Signal detection
MSB shift register
Laser velocimeter
AGC amplifier
ADC
Filter bank
Signal processor
CPU
LSB shift register
AGC

Figure 3. Block diagram of proposed frequency domain signal processor.

ORIGINAL PAGE IS
OF POOR QUALITY

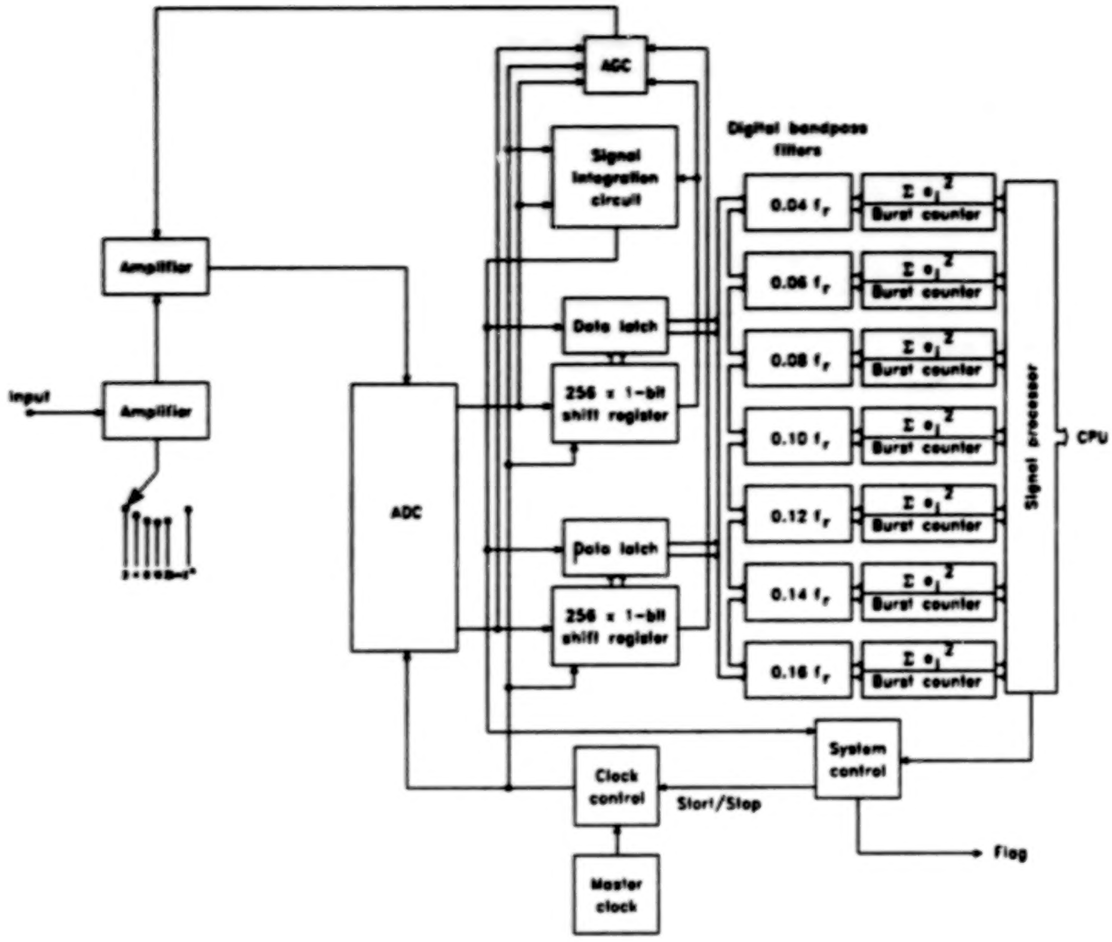
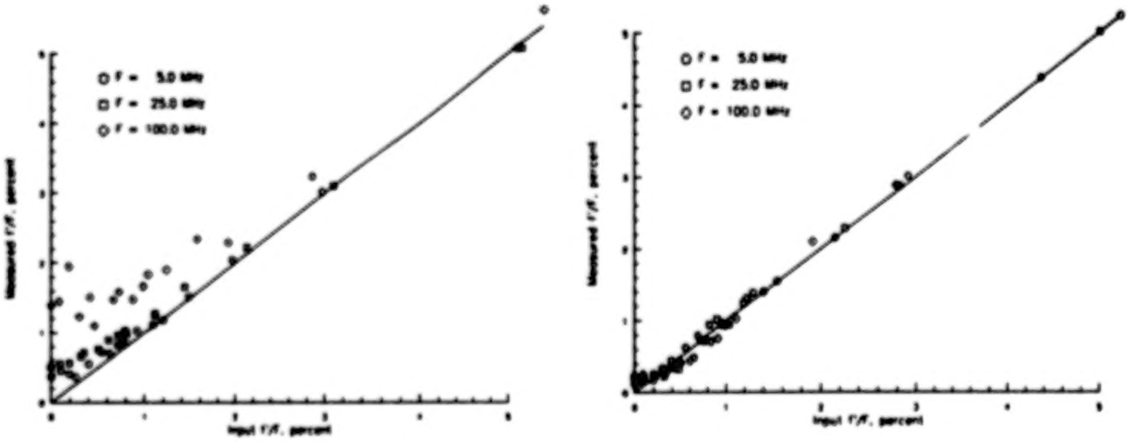

AGC
Signal integration circuit
Digital bandpass filters
Amplifier
Amplifier
Input
ADC
Data latch
256 x 1-bit shift register
Data latch
256 x 1-bit shift register
0.04 f_r
0.06 f_r
0.08 f_r
0.10 f_r
0.12 f_r
0.14 f_r
0.16 f_r
Σ e_i^2
Burst counter
Signal processor
CPU
System control
Clock control
Start/Stop
Master clock
Flag

Figure 4. Schematic diagram of proposed frequency domain signal processor.



(a) High-speed burst counter.

(b) Frequency domain processor.

Figure 5. Simulation comparisons of input f'/F to measured f'/F with mean oscillation frequencies of 5.0 MHz, 25.0 MHz, and 100.0 MHz.

ORIGINAL PAGE IS
OF POOR QUALITY

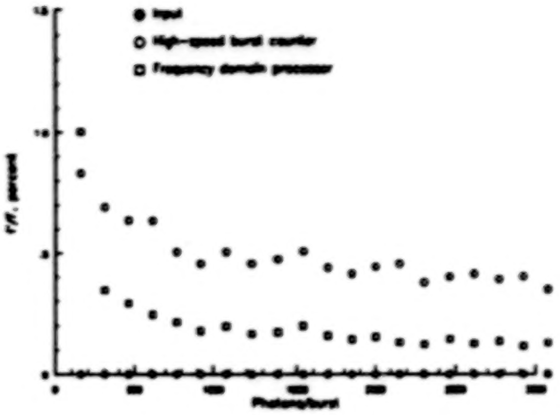


Figure 6. Simulation comparisons of input $f'/F = 0$ to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at mean oscillation frequency of 25.0 MHz.

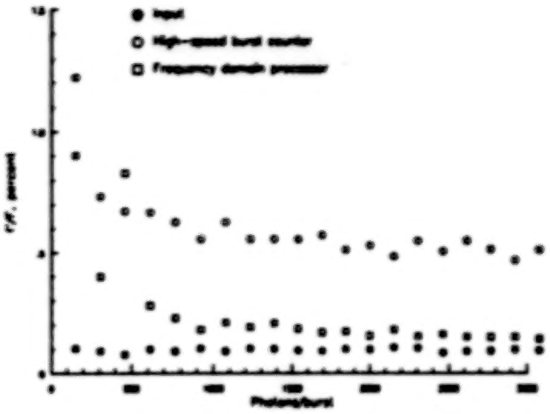


Figure7. Simulation comparisons of input $f'/F = 0.1$ to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise (photons/burst) at mean oscillation frequency of 25.0 MHz.

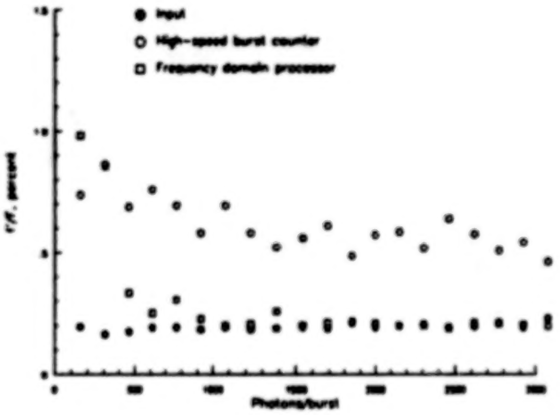


Figure 8. Simulation comparisons of input $f'/F = 0.2$ to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at mean oscillation frequency of 25.0 MHz.

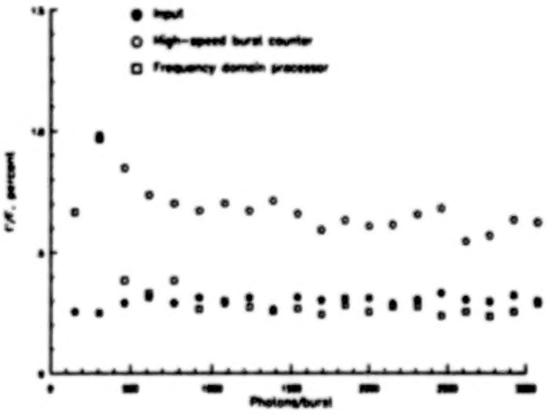


Figure 9. Simulation comparisons if input $f'/F = 0.3$ to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at mean oscillation frequency of 25.0 MHz.

ORIGINAL PAGE IS
OF POOR QUALITY

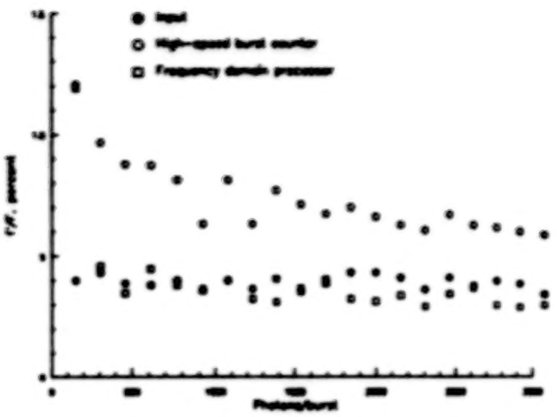


Figure 10. Simulation comparisons of input f'/F = 0.4 to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at mean oscillation frequency of 25.0 MHz.

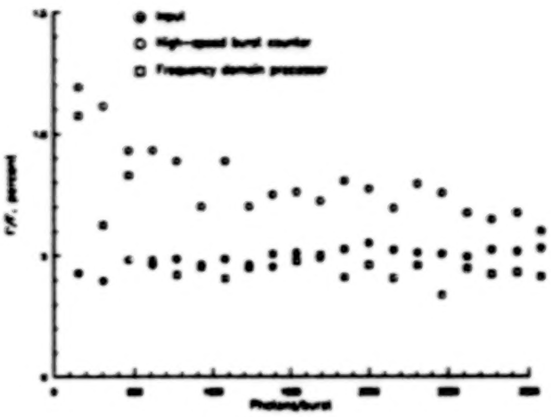


Figure 11. Simulation comparisons of input f'/F = 0.5 to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at a mean oscillation frequency of 25 MHz.

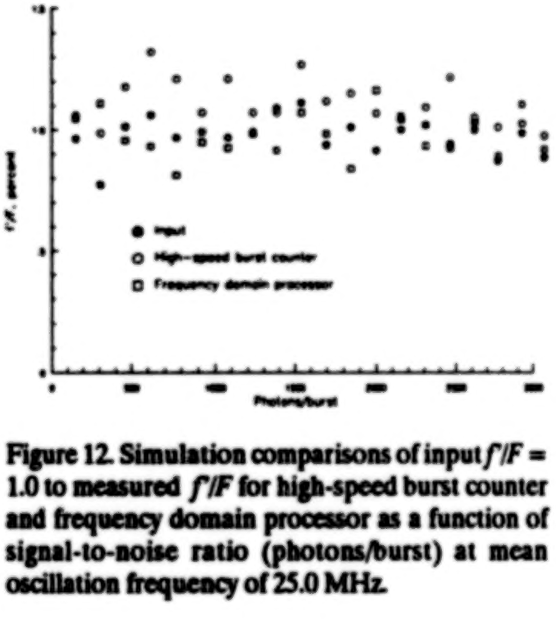


Figure 12. Simulation comparisons of input f'/F = 1.0 to measured f'/F for high-speed burst counter and frequency domain processor as a function of signal-to-noise ratio (photons/burst) at mean oscillation frequency of 25.0 MHz.

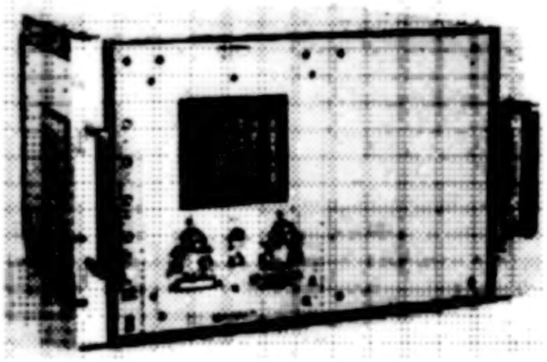


Figure 13. The FDP-3100 frequency domain processor.

ORIGINAL PAGE IS
OF POOR QUALITY

N91

24078

UNCLAS

N91-24078

A FIELD-DEPLOYABLE DIGITAL ACOUSTIC MEASUREMENT SYSTEM

David L. Gray, Kenneth D. Wright II,
and Wayne D. Rowland
NASA Langley Research Center
Hampton, Virginia 23665

ABSTRACT

A field-deployable digital acoustic measurement system has been developed to support acoustic research programs at the Langley Research Center. The system digitizes the acoustic inputs at the microphone which can be located up to 1000 feet from the van which houses the acquisition, storage and analysis equipment. Digitized data from up to 12 microphones is recorded on high density 8mm tape and is analyzed post-test by a microcomputer system. Synchronous and non-synchronous sampling is available with maximum sample rates of 12,500 and 40,000 samples per second, respectively. The high density tape storage system is capable of storing 5 gigabytes of data at transfer rates up to 1 megabyte per second. System overall dynamic range exceeds 83 dB.

INTRODUCTION

Researchers at the National Aeronautics and Space Administration, Langley Research Center, are engaged in acoustic research directed toward understanding and reducing noise generated by aerospace vehicles. The research encompasses both theoretical and experimental studies and involves numerous field tests and measurements of noise signatures from various types of vehicles. This paper describes a new field-deployable, digital acoustic measurement system developed to provide an acoustic measurement capability in support of this research. The need to replace the analog acoustic measurement systems, which have been used very successfully for many years, with digital systems was driven by several factors: first, the reduced availability of test vehicles for extended test periods and the limited life cycle of some components of many of the research vehicles being tested have made it essential that test time be reduced to a minimum; second, there is a need to reduce the time between completion of field tests and publication of the results. These requirements are more easily realized if the data is digitized during acquisition and processed post test in the field. Field processing allows alteration of the test plan after each day's testing, if needed, to fill out the test matrix and eliminates redundant tests that might be conducted if results were not available. Currently, Langley has four field instrumentation vans: three with analog instrumentation and one with digital instrumentation. Each van has acquisition, monitoring, and recording equipment for 12 microphone channels. Two new vans with digital instrumentation are being built. In addition to normal acoustic data, test vehicle position is measured using tracking instrumentation, local atmospheric conditions are acquired from a weather station set up at the test site, and earth acoustic impedance is measured. With the addition of the new digital system, all of the data including the non-acoustic data is in digital form which allows the integration of all data elements to be easily accomplished.

SYSTEM OVERVIEW

Figure 1 is a block diagram of the digital acoustic instrumentation system. The system is composed of acquisition and recording systems and a basic monitoring system with oscilloscopes, spectrum analyzer, meters, strip chart recorder, etc. The acquisition system has four elements: the remote digitizers located at the microphones to digitize the microphone analog outputs, the encoder/decoder subsystem, the display and control subsystem, and the digital acquisition-to-tape interface. All of these elements except the microphones and remote digitizers are in an instrumentation van that can be located up to 1000 feet away. The instrumentation vans are positioned at large distances from the microphones to reduce noise pickup from the van power generators and to avoid interference with the measurement. Using the analog systems, the long transmission lines from the microphones to the instrumentation van have created problems at times because the long lines act as antennas and pick up radio frequency signals which are demodulated by the analog instrumentation. The frequency of the demodulated signal falls in the audio band and appears as interfering noise to the acoustic measurement, in effect reducing the dynamic range of the measuring system. Since the digital system digitizes the data at the microphone and only digital data is transmitted on the transmission lines, the effects of the demodulated radio interference are negligible. Digitizing the data at the microphone has allowed a significant increase in the dynamic range (30-40dB) of the measurement, and this increased dynamic range has all but eliminated the need for operator gain changes to maintain a good signal-to-noise ratio. The other major system element is the recording system which is also located in the instrumentation van and is composed of a data conversion/buffering subsystem, the tape drive subsystem, and a digital-to-analog conversion subsystem. The digital recording system was developed as a low-cost alternative to the traditional digital instrumentation recorder, primarily because of the numbers of recorders required and the cost of the recorders. The new data recorders are relatively low cost and provide a significant new capability that is not available commercially. In the following paragraphs, the acquisition and recording systems will be discussed separately and in detail.

ACQUISITION SYSTEM

The acquisition system can be operated in either a synchronous mode, where the output of each microphone is sampled at the same instant of time, or in a nonsynchronous mode where the sample rate is the same for all microphones but there is no time correlation between samples. In the synchronous mode, sample time is controlled from the instrumentation van; and in the nonsynchronous mode, sample time is controlled by circuitry in each remote digitizer. The major elements of the acquisition system are the remote digitizers, the encode/decode subsystem, the display and control subsystem, and the digital-to-analog converter subsystem. Each of these elements will be discussed.

Remote Digitizers

The remote digitizer is shown in block diagram form in figure 2. This remote electronic system is powered by two 6 volt sealed lead acid batteries driving a DC to DC converter which supplies several voltages to the electronics. The microphone is a condenser type with an integral preamplifier. The output of the microphone preamplifier is input to a digitally controlled variable gain amplifier with gains of 1 to 128. The output of the amplifier is low-pass filtered and sampled either in a synchronous mode with the sample time controlled remotely from the instrumentation van or nonsynchronously, where the sample time is controlled in each remote digitizer by a precision clock in the logic control circuitry. The sampled analog data is converted to digital nonreturn to zero (NRZ), offset binary format by the analog-to-digital converter. The digital data is then converted to a serial data stream, encoded in a Manchester II code and transmitted to the instrumentation van. The Logic Control Circuitry controls the entire process.

Encode/Decode System

The encode/decode subsystem (figure 1) is located in the instrumentation van. When in the synchronous data mode, the system encodes the gains to be set on the variable gain amplifier in the remote digitizer in a Manchester code which is sent to each remote digitizer each time it is commanded to sample the microphone data. Communications between the instrumentation van and the remote digitizer can be half-duplex or full-duplex. Only half-duplex has been used in the system to date because of the lighter, smaller cable requirements. The subsystem also decodes the Manchester coded data received from the remote digitizer in both the synchronous and nonsynchronous modes of operation.

Display and Control System

The display and control subsystem (figure 1) is used to display the amplitude of the data from each channel, set the gain of variable gain amplifier in the remote digitizer, detect and alert the operator of overload conditions in the remote digitizer, and alert the operator of any faults in the communication of information and data between the instrumentation van and the remote digitizer. The display and control chassis is pictured in figure 3.

Digital-to-Analog Conversion System

Each data channel has a digital-to analog converter to allow the operator to monitor the data from each microphone. The D/A converter for each channel monitors the data between the encode/decode system and the acquisition/tape interface. Converting the data to analog at several points in the serial path of the acquisition and recording system allows the operator to isolate system electronic problems to specific elements of the system.

Acquisition to Tape Interface

The 12 microphone data channels and 8 additional channels of information, i. e., time code, run number, voice annotation, etc., are put into the proper format and clocked into the recording system by this interface. There are 20 parallel channels for transfer of NRZ serial data. Each channel is composed of three signals: data, clock and strobe.

Acquisition System Specifications

The following table lists the most important features and specifications for the acquisition system.

Table 1. Acquisition System Features and Specifications

Sample Modes: Synchronous and Non-synchronous.

Sample Rates: Half-Duplex, synchronous, 12,500 samples/second maximum. Half-Duplex, non-synchronous, 40,000 samples/second maximum. Full-Duplex, synchronous and non-synchronous, 40,000 sample/second maximum.

Digitization: 16 Bit, Offset Binary.

Gain Range: 102-142 dB (128 to 1). Gain controlled from instrumentation van in synchronous mode of operation.

Cable Length: 2000 feet with bit rate of 330,000 bits/second. 1000 feet with bit rate of 1,000,000 bits/second.

System Clock: Variable, 4×10^6 to 12×10^6.

Cable Type: Half-Duplex, single twisted, shielded pair. Full-Duplex, two twisted, shielded pairs.

Dynamic Range: >83 dB.

Bandwidth: Two fixed bandwidth, 8-Pole Butterworth low pass filters, 1000 Hertz cut-off and 10,000 Hertz cut-off.

Data Encoding: Manchester II code.

Battery Life: 20 hours with 1-2 hours reserve before recharging.

Auxiliary Signal Input: +/-10 volts.

LED Bar Display: Displays amplitude with peak hold and incorporates LED to indicate loss of Manchester signal from remote digitizer.

Modular Construction: Construction modularized to allow upgrades such as improved A/D converter and allow the system to be used for other data acquisition application.

Control: Manual or computer control of gains and sample rate. Computer to provide a data monitoring capability in the future.

Data Monitoring: D/A converter for each data channel.

Data Channels: 12.

RECORDING SYSTEM

The recording system uses 8mm helican scan, cartridge tape drives to store 20 channels of acoustic data and test information on a single track on the recorder. Figure 4 is a photograph of the tape drive system. One tape drive is used to record the primary data tape, and the second drive is used to simultaneously record a backup tape. The second drive also provides redundancy in case of drive or electronic failure. Each tape drive has its own microprocessor controller and electronics. Data compression systems are combined with the tape drives to provide several possible configurations and data transfer rates. Although all of the various configurations have not been tested, the expected transfer rates for some of the configurations with and without data compression are given in table 2. The recording system includes the data conversion and buffering subsystem, the tape drive subsystem, and the digital-to-analog conversion subsystem. Each of these subsystems will be described.

Data Conversion and Buffering Subsystem

This subsystem accepts 20 parallel channels of serial data coming from the acquisition system and converts the data in each channel to a parallel format. The parallel 16-bit-wide data samples for each channel are loaded into 2000 byte First In First Out (FIFO) speed matching buffers. The data in the FIFO's is output to the tape drive subsystem. Also included in this subsystem is a digital-to-analog converter that can be switched to any data channel. This provides a data monitoring as well as troubleshooting capability.

Tape Drive Subsystem

The tape drive subsystem, pictured in figure 4, is composed of 8mm tape drives, microprocessor controllers, Small Computer Systems Interface (SCSI) controller, logic to activate the front panel switches controlled by the microprocessor, run number switch, logic to detect loss of a channel or data dropout, and logic and display of the time remaining on the tape. In the record and playback modes, the microprocessor is used to activate the front panel switches in a set sequence to reduce the possibility of pressing the wrong switch. When a tape is loaded, it is checked to see if it has been used. If data is present, the system will only allow playback of the tape. The 8mm drives use the Small Computer System Interface (SCSI) and the electronics are included in each drive. The tape drive subsystem electronics interface to the SCSI bus through a SCSI integrated circuit controlled by the microprocessors. In the record mode, the microprocessor initiates data transfer from the FIFO's to the microprocessor bus; but once started, the SCSI integrated circuit continues data transfer from the microprocessor bus to the SCSI bus in a Direct Memory Access (DMA) mode until the stop switch is activated. Transfer rates up to 2 megabytes/second are possible with the SCSI controller and electronics used although the tape drives will not accept data at this rate. The tape subsystem has a playback mode primarily for setup and troubleshooting. During playback the data is demultiplexed and input into digital-to-analog converters. Data compressors not shown on the block diagram in figure 1 are connected between the SCSI bus output of the SCSI integrated circuit and the SCSI bus input to each tape drive. Since the data compressor resides on the SCSI bus, it is not shown as a separate subsystem.

D/A Conversion Subsystem

The digital-to-analog conversion subsystem is used to monitor the digital data going onto the tapes during the record mode and coming off the tapes during the playback mode. Since the 8mm drives do not provide a true read-after-write capability, the data is monitored at the last available point before being recorded. The tape drives perform internal read-after-write to do error correction as the data is recorded. This provides a very low bit error rate for the recorded data. Reliability is provided by using redundant tape drives, monitoring of the data on the SCSI bus, and error correction.

Table 2. Recording System Features and Specifications

Tape Drives: 8mm Helican Scan, cartridge tape.

Transfer Rates: Uncompressed data, 250 kilobytes/second. With upgraded tape drives, 500 kilobytes/second. Compressed data and upgraded tape drives, 2 megabytes/second.

Storage Capability: Upgraded drives, no compression, 5 megabytes. Upgraded drives with data compression, minimum of 10 megabytes.

Redundancy: Dual tape drives and electronics.

Control: Each tape drive is controlled by a separate 68000 microprocessor. The front panel switches are controlled to prevent nonallowable switch sequences. Recorded tapes are automatically prevented from being overwritten. Loss of a data channel during acquisition is automatically detected, an LED is turned on, and saturation levels are stored on tape for the failed channel.

Housekeeping Data: Gains, run number, time code, and voice annotation.

Multiplexed Channels: 20.

Data Monitoring: D/A converter for each data channel.

Bus: Small Computer Systems Interface (SCSI).

Display: Time remaining on the tape, channels deactivated or failed, and data tape loaded.

Data Buffering: 2 kilobyte FIFO buffer on each channel.

Data Interface: Serial or parallel.

CONCLUSION

A field deployable, digital acquisition and recording system has been developed that is very versatile and although specifically developed to satisfy the requirements for an acoustic measurement system, can be applied to other measurement needs. The system was designed to allow improved components to be incorporated as they become available and are needed, and also to allow the acquisition and recording systems to be used independently of one another or with other systems. The acquisition system has operated very successfully in several field tests. The recording system is currently being integrated with the acquisition system and being tested in the laboratory. The first field test of the complete acquisition and recording system is scheduled for January 1991. Future additions to the system will include computer monitoring of data channel, automatic activity logging, and expanded tape annotation.

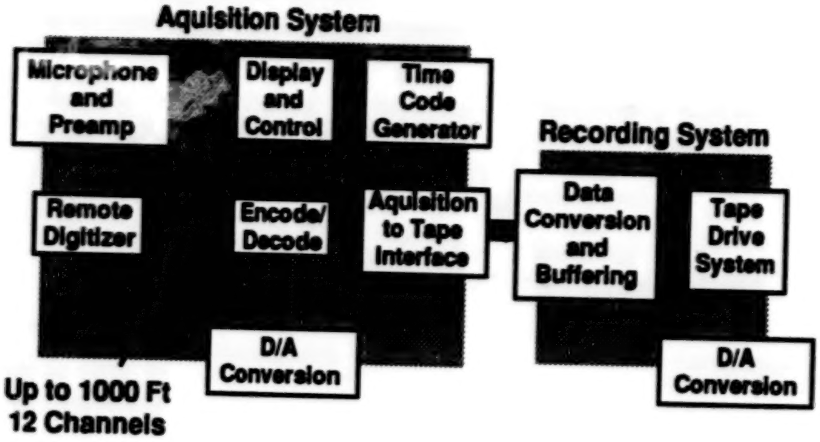

Aquisition System
Microphone and Preamp
Display and Control
Time Code Generator
Recording System
Remote Digitizer
Encode/ Decode
Aquisition to Tape Interface
Data Conversion and Buffering
Tape Drive System
D/A Conversion
D/A Conversion
Up to 1000 Ft
12 Channels

Figure 1: Digital Acoustic Measurement System Block Diagram

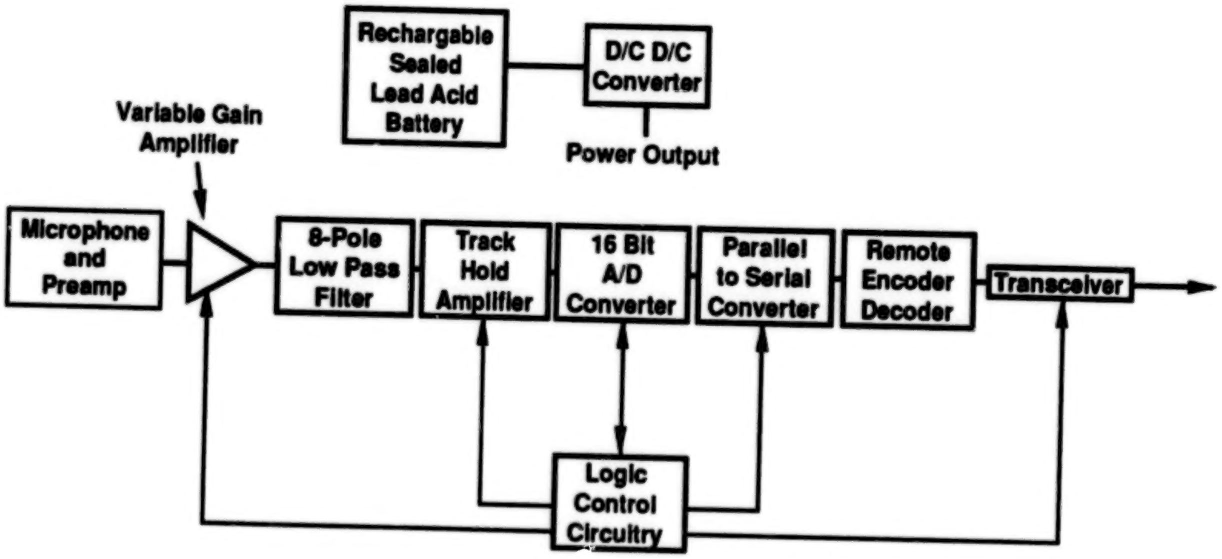

Rechargable Sealed Lead Acid Battery
D/C D/C Converter
Power Output
Variable Gain Amplifier
Microphone and Preamp
8-Pole Low Pass Filter
Track Hold Amplifier
16 Bit A/D Converter
Parallel to Serial Converter
Remote Encoder Decoder
Transceiver
Logic Control Circuitry

Figure 2: Remote Digitizer Block Diagram

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

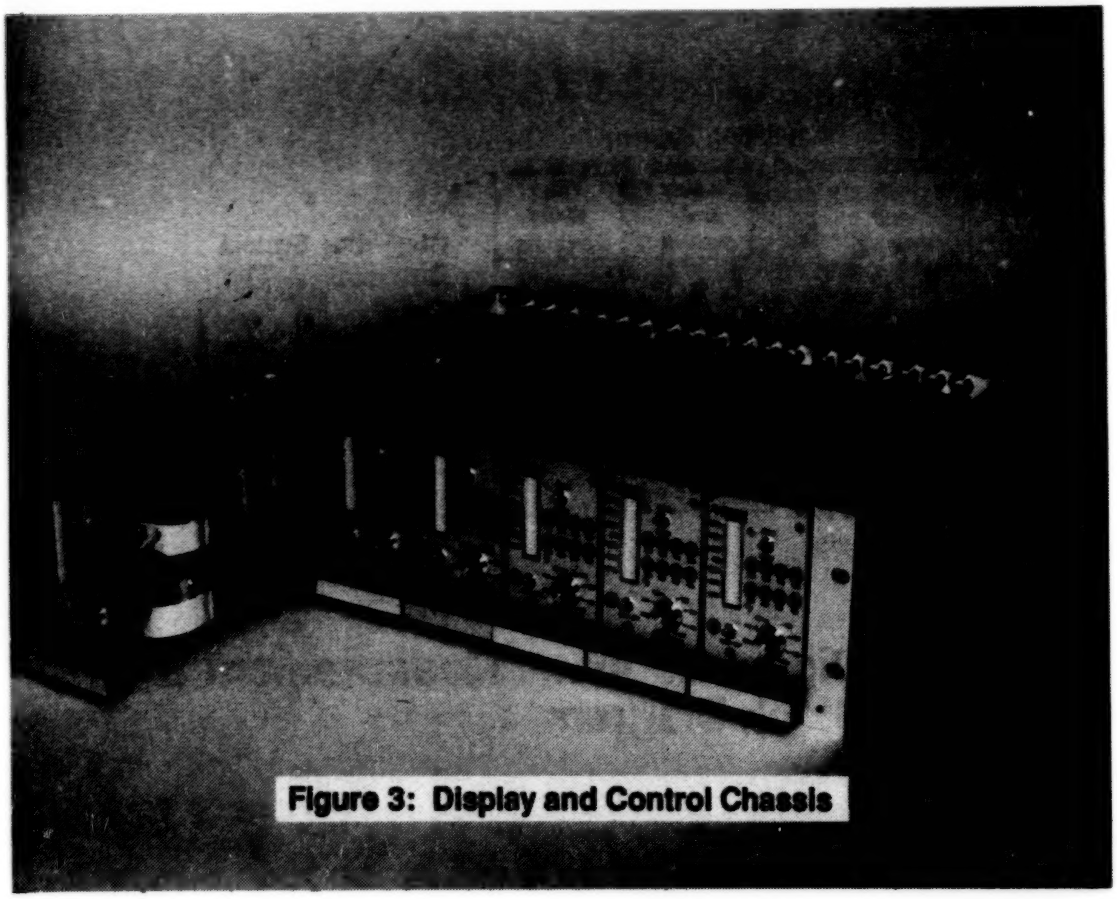


Figure 3: Display and Control Chassis

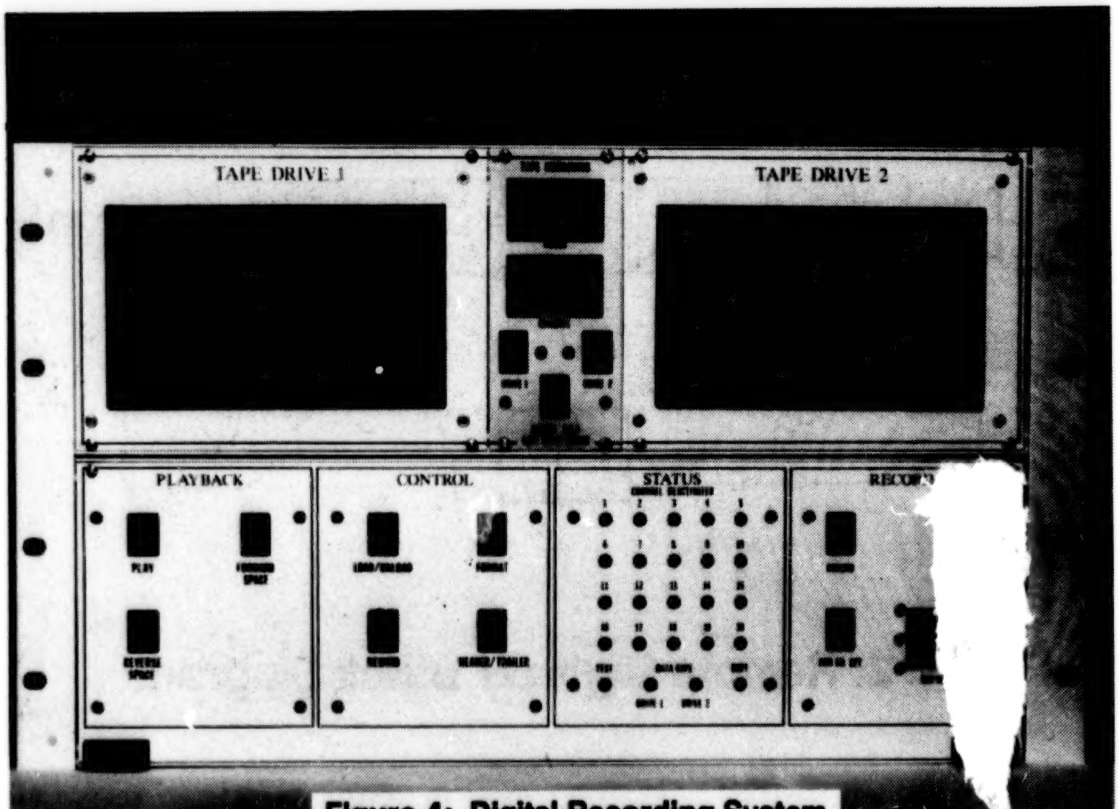

TAPE DRIVE 1
TAPE DRIVE 2
PLAYBACK
PLAY
FORWARD SPACE
REVERSE SPACE
CONTROL
LOAD/UNLOAD
FORMAT
REWIND
HEADER/TRAILER
STATUS
1
2
3
4
5
RECO

Figure 4: Digital Recording System

N91

24079

UNCLAS

N91-24079

LASER OPTICAL DISK POSITION ENCODER WITH ACTIVE HEADS

Eric P. Osborne
Electromechanical Branch
Goddard Space Flight Center

ABSTRACT

An angular position encoder is discussed that minimizes the effects of eccentricity and other misalignments between the disk and the read stations by employing heads with beam steering optics that actively track the disk in directions along the disk radius and normal to its surface. The device adapts features prevalent in optical disk technology toward the application of angular position sensing.

INTRODUCTION

In the following discussion, brief descriptions of both conventional encoder technology and optical disk technology are provided. A hybrid encoder is then considered which combines these two technologies. Apart from higher resolution, numerous other advantages are foreseen for this hybrid device.

CONVENTIONAL OPTICAL ENCODERS

Physical Description and Operation

A conventional rotary optical encoder consists of the elements depicted in Figure 1. Photodetectors are used to sense the motion of a disk which has been affixed to a shaft whose orientation is to be monitored. Encoded on the surface of this transparent disk is a series of dark lines running in the radial direction. These lines form a circular track bounded by an inner and outer radius. A light source is placed on the opposite side of the disk. As the shaft rotates, the photo detectors sense the changing illumination as a series of pulses. In order to collimate the light, a mask and a pair of lenses are generally employed.

The encoder just described is of the incremental type. It simply counts pulses from some reference orientation and computes the angular rotation in proportion to the number of counts registered. Multiple tracks or read stations placed out of phase with respect to one another are used to sense direction of rotation. Phasing of read stations also generates phased wave forms which can be interpolated electronically to yield higher resolutions.

Error Sources

Several factors affecting the performance of encoders have already been alluded to. Eccentric mounting of the disk will result in a cyclical variation in the location of the track with respect to the read station. The radius at which counts are being taken varies accordingly. Since the angular position is calculated by assuming a fixed radial distance in combination with the observed arc length, an angular error results.

Similar behavior results when there is radial play at the interface between the shaft and the bearing bore. Since the shaft, and the affixed disk, are free to wander, the radial position of the track relative to the read station varies in an unpredictable fashion. Once again angular errors result.

Axial play can cause problems as well. Because of the optical arrangement typically used, the distance between the disk and the emitter is critical, particularly in high resolution devices. Variations in this distance cause the illumination passing through the disk to vary even when the disk is not rotating. These variations can confuse the counter. Disk warp can introduce the same kinds of effects.

Two other parameters which encoders are sensitive to are disk to shaft perpendicularity and the circularity of the track on the disk. These and the aforementioned error sources are strictly mechanical in nature. Errors can be introduced by thermal or electrical effects as well, but they will not be addressed here. Multiple read stations evenly spaced around the disk are often used in the higher resolution devices to allow averaging out of disk wobble and warp.

Encoder Constructions

Encoders are available as either pre-assembled units or modular kits. Pre-assembled units contain all components internally. Each unit includes its own shaft to which the disk is affixed internally. At installation, the shaft of interest is attached to the encoder shaft via a shaft coupling, and the body of the encoder is bolted to the shaft housing.

Modular encoders consist of two independent pieces: the disk, and the encoder body which includes the optics. In this case, the disk is coupled to the subject shaft independent of the encoder body. The encoder body is then placed over top of the disk. Generally some provision is provided for adjusting the relative position of the encoder body with respect to the disk. The ability to align these components relative to one another directly influences the accuracy and the resolution achievable.

The advantage of the preassembled units is that the alignment of the optics with respect to disk is maintained by a bearing internal to the encoder. The manufacturer has the tools to ensure that this alignment meets tight tolerances. The alignment of the encoder shaft to the subject shaft, however, is generally the responsibility of the user. Care must be taken not only to align the encoder shaft to the subject shaft properly, but to ensure that no significant loads are passed through the encoder bearings due to relative motion of the two shafts. Such loads may introduce misalignment, or worse yet, lead to deformation of the disk and other anomalies. Flexible couplings have been developed to address this problem.

High resolution encoders have smaller optical paths which are more easily fouled by contaminants. It is impractical to allow assembly of these encoders by the end user in an uncontrolled environment. Pre-assembled units come sealed from the factory and are more resistant to contaminants than the modular types.

Modular encoders offer the advantage of being relatively inexpensive, and insensitive to external loads because they employ no internal bearings. Since they employ no additional bearings, one may argue that they are inherently more reliable. They also require no internal lubrication which is an important plus for applications in clean or low pressure environments. The chief disadvantage of these encoders is that the alignment of the encoder body relative to the disk is the user's responsibility. Accordingly, modular encoders are not used in high resolution applications [1].

OPTICAL DISK TECHNOLOGY

Optical disk technology involves the storage and retrieval of data from a disk through the use of lasers. An embodiment of this technology with which most persons are familiar is Compact Disc, or CD, technology. Drives which handle optical disks are currently being sold as computer peripherals. The key selling point of these devices with respect to competing technologies in the mass storage market is the high density to which information can be stored. On a standard CD one bit occupies a spot only 0.5 microns (0.00002") in diameter. These bits, or pits, are burned into with a low power laser while the disk is rotating. This "burning in" of data constitutes the process of recording. The laser is pulsed appropriately to lay a stream of pits down in either a spiral track or a series of concentric tracks. Once the pattern has been burned into the disk, its entire surface is coated with a thin metallic deposition. Thus, the bottoms of the pits as well as the regions between pits (referred to as lands) have nearly the same reflectivity.

The principles of interferometry are applied in order to read the disk. The power to the laser is reduced such that it acts only as an illuminator. As the disk rotates, the light from the laser is focused via a movable objective lens onto a particular track. Variations in reflectivity occur at the transitions between pits and lands as a result of constructive interference, and these variations are sensed by the photodetector. Figure 2 depicts the arrangement of components within a typical read station.

In order to keep the beam focused to the required degree, the read station used in an optical disk drive is equipped with servo controlled actuators. One of these moves the objective lens of the read head vertically to control focus and thereby accommodate disk warp. In order to accommodate disk wobble, a second actuator moves the objective laterally to maintain view of the track of interest at all times.

The feedback required to achieve this closed loop actuation is provided by the photodetector. The photodetector typically consists of a four element photodiode array. An amorphic objective lens is also used. When the beam is properly focused, the reflected light forms a circle on the array, equally illuminating all elements. As the disk moves out of focus, the beam rotates and deforms into an elliptical shape. This results in an imbalance in the illumination of the array elements. If the disk is too far away, the ellipse will form along one diagonal, and if it is too close it will form along the other.

Tracking information may be obtained by monitoring the signals on adjacent elements in the photodetector array. As the beam drifts off track it encounters the transition between pit and land. The resulting constructive interference causes a decrease in illumination over some of the detector array elements. The imbalance in the signals received from the elements is used to generate an appropriate servo signal.

More sophisticated tracking approaches are often used. In the three beam method, two beams are split off the main beam by using a diffraction grating. The grating is rotated out of plane such that these flanking beams strike the disk on opposite sides of the track and fore and aft of the main beam. Two additional elements are incorporated into the photodetector array to monitor the intensity of the flanking beams. Once again, appropriate servo signals are generated by comparing array element illumination signals.

The active focus and tracking actuators not only accommodate warp and wobble, but also make the device insensitive to vibration. The actuators are simple voice coil motors in which the objective is attached to a movable bobbin. This bobbin reacts to magnetic field variations in the surrounding control coils. It is physically connected to the surrounding structure by a spring metal diaphragm. There are no wear surfaces within the actuator.

Concerns about dust and contaminants blocking the field of view of the laser are resolved in the following manner. The information layer is actually imbedded in the disk as indicated in Figure 3. A lacquer coating protects this information layer. The optics within the read head are designed such that the output beam strikes the disk surface over a relatively large area about 0.7 mm (0.027") in diameter. As a result of the lacquer's relative index of refraction, the beam converges rapidly inside the disk to a spot only 1.0 micron in diameter. Thus, a particle on the surface would have to be quite large to interrupt the beam. In order to handle the eventuality of beam blockage, the data encoded on an optical disk is usually laid out such that the input data can be reconstructed despite the loss of a bit [2].

A HYBRID ENCODER

Physical Description

At NASA's Goddard Space Flight Center, a hybrid encoder is being developed which adapts the advantages of optical disk technology toward the problem of shaft orientation sensing. This hybrid encoder employs the optical approach and disk construction of optical disk technology. Potential applications include high resolution scanning instruments which require long life such as those to be employed on the EOS Platform. An isometric view of the hybrid encoder is provided in Figure 4. The read stations used are identical to those found in optical disk drives. The tracking strategy and the disk pattern employed, however, are designed to address the particular concerns of angular position sensing.

In an optical disk, the proximity of the pits to one another is limited by manufacturing process considerations. There are regions between pits of greater span than the diameter of the light beam. As the main beam passes through these regions, the illumination on all photodetector array elements is the same and no useful feedback is provided from which to derive radial position. For a constantly spinning disk, this is not a problem as a pit soon comes into view,, and the necessary feedback is made available. However, difficulty is encountered if it is desired to maintain head to disk alignment when the disk is stationary. There is a good probability that the disk will be oriented such that there are no pits within the field of view of the beam. Since no feedback is provided, motion of the disk cannot be observed. This is untenable in the encoder application because the very purpose of the device is to detect motion.

In order to alleviate this problem, a special disk pattern has been proposed for use in conjunction with the three beam tracking method. The pits are sized and located such that a pit is always visible in either the main beam or the flanking beams. Track-to-track spacing and the orientation of the diffraction grating are selected such that, should the main beam be off track, at least one of the flanking beams intercepts a pit regardless of the relative alignment of pits between tracks.

In order to allow for acquisition of a common track by beams from multiple heads, regardless of their initial radial positions, the arrangement of tracks depicted in Figures 5a and 5b is used. Two guard bands define a region of data tracks consisting of pits spaced at even intervals. The number of intervals is fewest near the guard bands and greatest along a central target track.

Operation

At setup the heads are aligned over the region defined by the guard bands. The disk is rotated at a constant rate,, and the servos are activated for each head. Each head acquires the nearest track and identifies the frequency of the signal fluctuations received. The beam is then moved over one track pitch, and the frequency associated with the adjacent track is identified. By comparing frequencies, the read station determines whether it is moving toward or away from the track having the highest frequency (the target track). Once the servos have locked onto the target track, the track radius is fixed, known and identical for all read stations. Since the read station moves with the track, the radius upon which angular calculations are based does not vary as the disk wobbles, as it does for a conventional encoder with fixed read stations.

Once the device has been setup as described, it is used just as a conventional incremental encoder. The variations in the detector array signals coming from each head are averaged, counted, and interpreted as angular rotations.

Advantages Of The Device

The hybrid encoder is modular in construction such that the disk and the encoder body have no direct physical connection. This results in an encoder which is tolerant of external loads. Damage at installation or during operation is therefore less likely than it is for a high resolution conventional encoder. Because the device employs no internal bearings, the reliability of the device is enhanced. The absence of these bearings also eliminates a potential source of contaminants either in the form of lubricants or wear debris.

Because the hybrid encoder uses read stations which employ internal servo controlled actuators, the encoder is actively self-aligning. Thus, warp and wobble in the disk are accommodated. External vibration, which might introduce misalignment between the encoder body and the disk, is also accommodated.

The disk used in the hybrid encoder is of the type used in optical disk drives rather than the photolithographic type used in conventional optical encoders. Thus, two important features are gained. First, because a laser is used to encode the disk, smaller markings may be encoded on the disk. A higher resolution is therefore obtainable for an encoder of a given diameter. As recording techniques and materials are refined in the various types of mass media, this encoder concept may be adapted to the media offering the greatest bit density. So even higher resolutions may be feasible.

The second advantage afforded by using the optical disk along with the associated optics is that the disk is more tolerant of contamination and physical abuse. This is due to the convergent optical path and the fact that the information layer is within the disk rather than on the surface.

The disk used in the hybrid encoder has a band of concentric tracks, any of which the read station may acquire. Alignment to the target track is achieved automatically using available feedback. Because of this,, the effort required to perform initial alignment and setup of the encoder is greatly reduced [3].

CONCLUDING REMARKS

This device shows promise, but requires an investment in time and resources. Hardware experimentation has begun, although it has been on hold for some time. Limited error sensitivity analyses have also been performed, but further study is required.

REFERENCES

1. Techniques for Digitizing Rotary and Linear Motion. Dynamics Research Corporation, 3rd ed., November 1980.

2. Watkinson, J.: The Compact Disc System. Electronics & Wireless World, January 1985, pp. 69-71.

3. Osborne, Eric P.: Laser Optical Disk Position Encoder with Active Heads. Patent Application. NASA Case No. GSC 13,175-1, 1990.

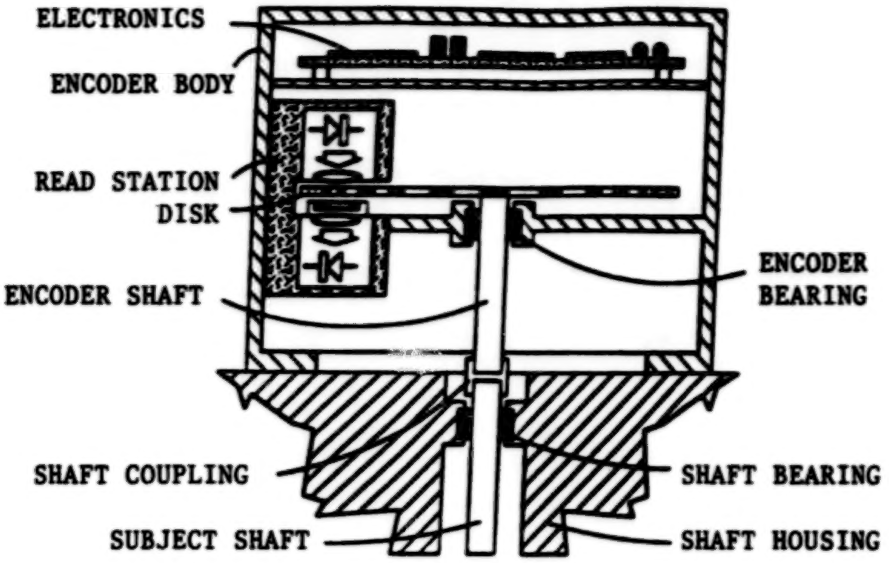

ELECTRONICS
ENCODER BODY
READ STATION
DISK
ENCODER SHAFT
ENCODER BEARING
SHAFT COUPLING
SHAFT BEARING
SUBJECT SHAFT
SHAFT HOUSING

FIGURE 1 - CONVENTIONAL ENCODER

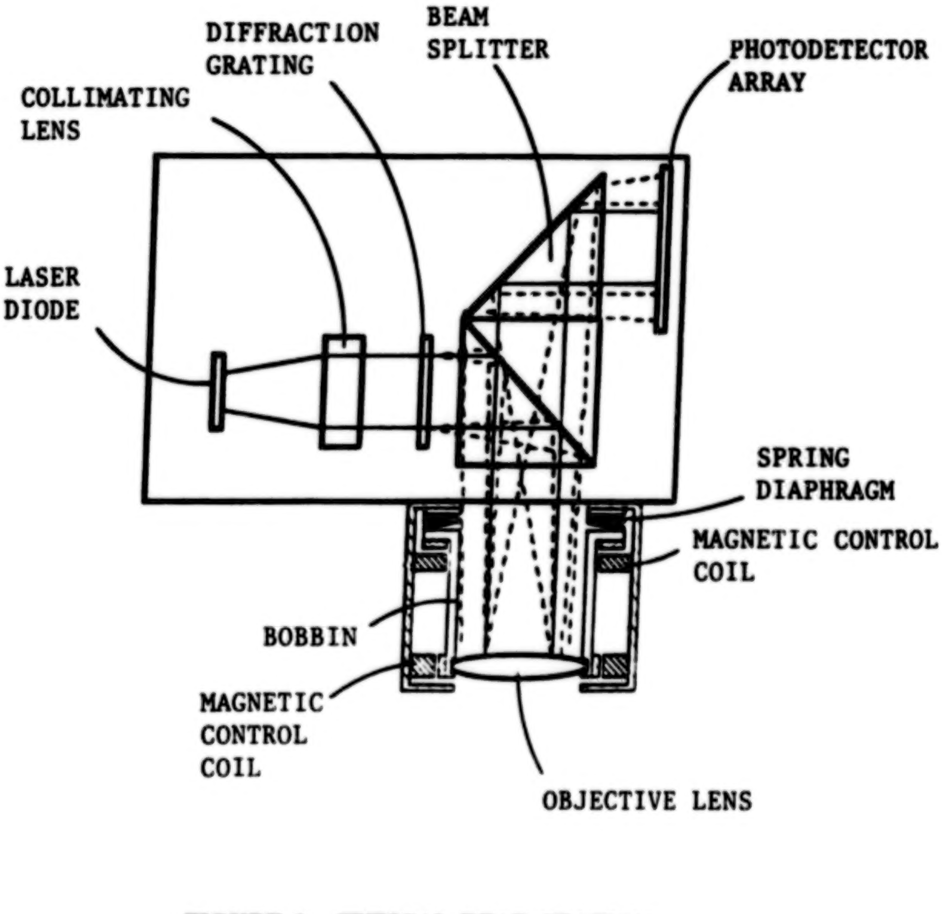

DIFFRACTION GRATING
BEAM SPLITTER
PHOTODETECTOR ARRAY
COLLIMATING LENS
LASER DIODE
SPRING DIAPHRAGM
MAGNETIC CONTROL COIL
BOBBIN
MAGNETIC CONTROL COIL
OBJECTIVE LENS

FIGURE 2 - TYPICAL READ STATION

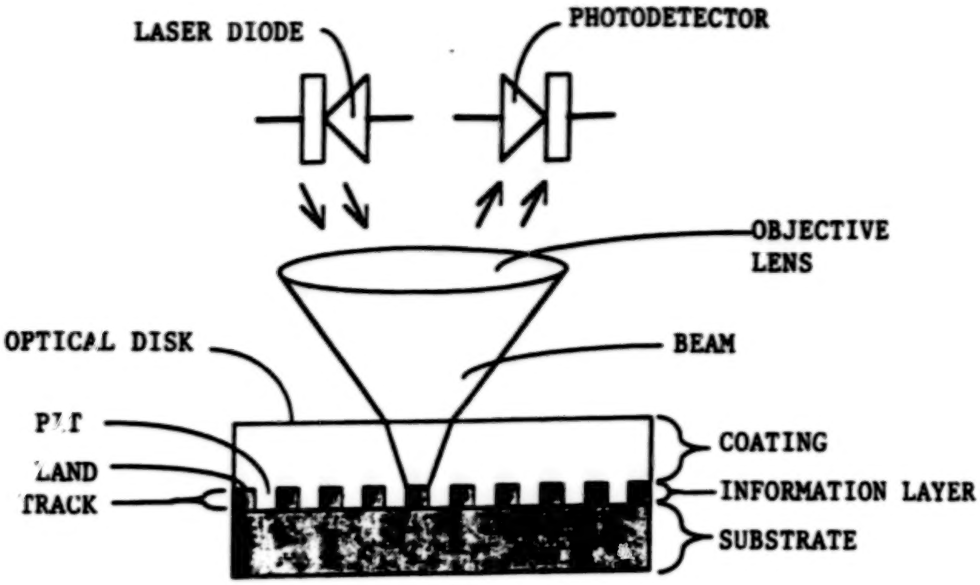

LASER DIODE
PHOTODETECTOR
OBJECTIVE LENS
OPTICAL DISK
BEAM
PIT
LAND
TRACK
COATING
INFORMATION LAYER
SUBSTRATE

FIGURE 3 - BEAM PATH SCHEMATIC

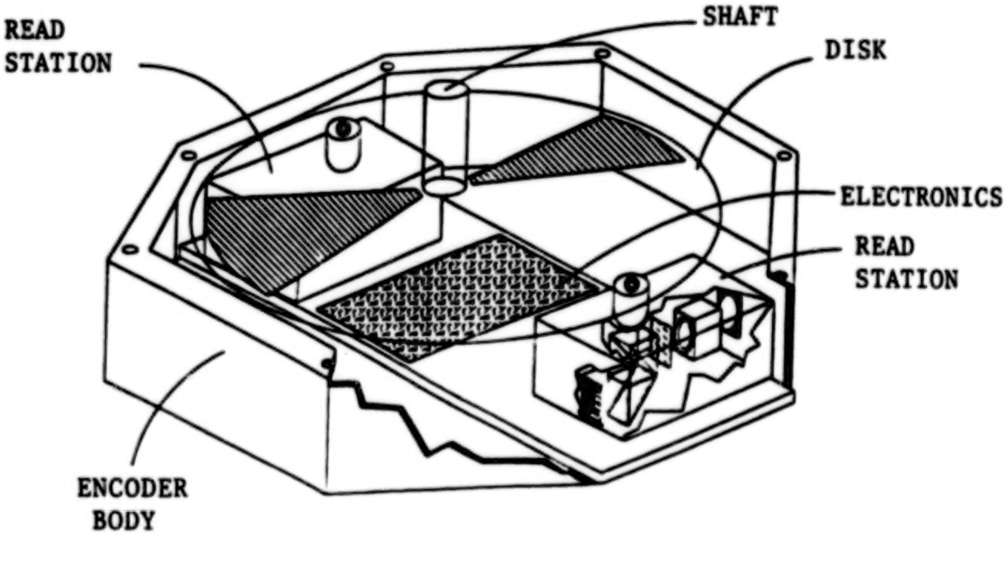

READ STATION
SHAFT
DISK
ELECTRONICS
READ STATION
ENCODER BODY

FIGURE 4 - ISOMETRIC VIEW OF HYBRID ENCODER

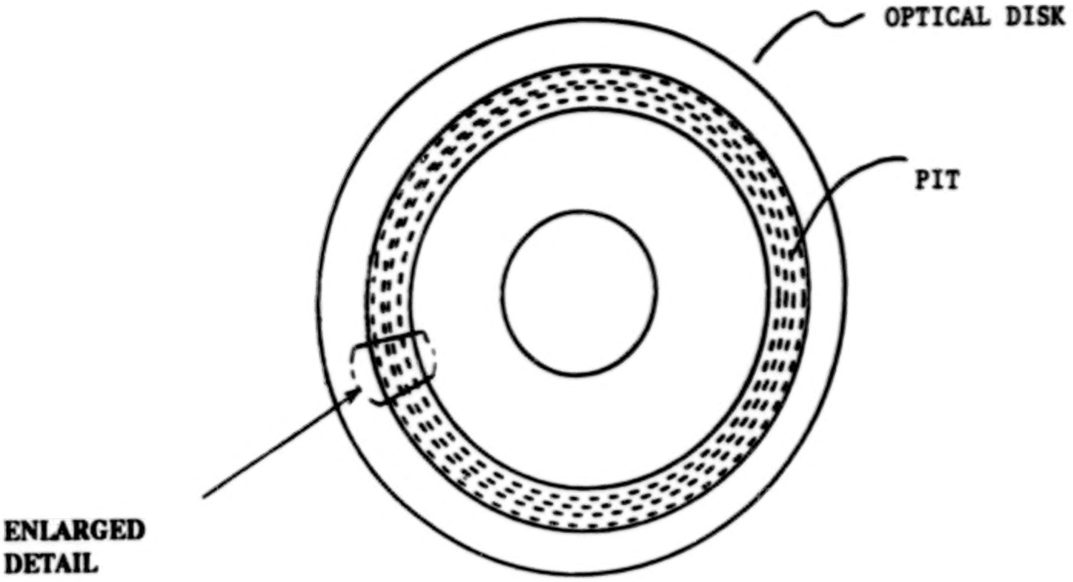
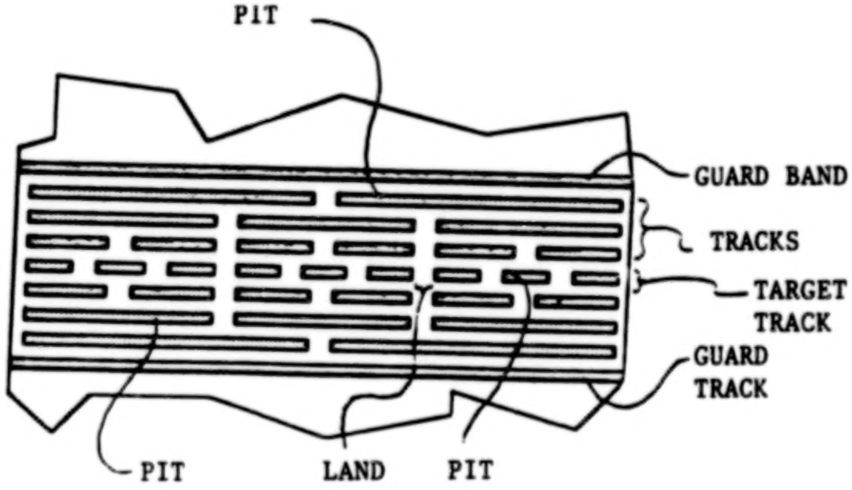

OPTICAL DISK
PIT
ENLARGED
DETAIL

FIGURE 5A


PIT
GUARD BAND
TRACKS
TARGET
TRACK
GUARD
TRACK
PIT
LAND
PIT

ENLARGED DETAIL

FIGURE 5B

FIGURE 5 - HYBRID ENCODER TRACK ARRANGEMENT

SESSION H - SUPERCONDUCTIVITY

Wednesday November 28, 1990

- **Applications Of High-Temperature Superconductors**
- **Superconducting Microwave Electronics At NASA Lewis**
- **Superconductive Wires And Devices For Cryogenic Applications**
- **Melt-Processed Bulk Superconductors Fabrication And Characterization For Power And Space Applications**
- **Large Gap Magnetic Suspension System Ground-Based Experiment**

N91

24080

UNCLAS

N91-24080

AEROSPACE APPLICATIONS OF HIGH TEMPERATURE SUPERCONDUCTIVITY

V. O. Heinen and D. J. Connolly
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, OH 44135

Abstract

Space application of high-temperature superconducting (HTS) materials may occur before most terrestrial applications because of the passive cooling possibilities in space and because of the economic feasibility of introducing an expensive new technology which has a significant system benefit in space. NASA Lewis Research Center has an ongoing program to develop space technology capitalizing on the potential benefit of HTS materials. The applications being pursued include space communications, power and propulsion systems, and magnetic bearings. In addition, NASA Lewis is pursuing materials research to improve the performance of HTS materials for space applications. This paper will discuss some of the applications we are pursuing. The HTS communications program is described in another paper in this proceedings.

Introduction

The discovery, in 1987, of ceramic materials which are superconducting at temperatures above 77 K initiated a new era in which superconducting technology could be envisioned as enabling a host of exciting new systems in the power, transportation, communications, and data processing industries. In the succeeding three years there has been steady progress in research to address concerns for the limitations of the new superconducting materials. The availability of thallium based materials which are superconducting up to 125 K has become firmly established. Yttrium based films with critical current densities (J_c) above $2X10^6$ amp/cm^2 at 77 K in magnetic fields of several Tesla, can now be routinely made by numerous laboratories world wide.

Impressive progress is also being made in the development of bulk superconducting materials with high critical current density - a parameter that is essential to power and propulsion applications. Various approaches to microstructural control have yielded J_c values near 10^7 A/cm^2 at 77 K and 1 Tesla.[1]

The application of superconducting technology has previously been limited by the requirement of cooling to near liquid helium temperatures. A superconducting transition temperature above 77 K and the cold background temperature in space reduce the cooling requirements significantly. The natural cooling ability of space as well as the acceptability, in space-based applications, of high up-front development costs suggest that space application will lead the way in exploiting the promise of high-temperature superconductivity (HTS).

The NASA Lewis Research Center is performing materials research to improve the performance of HTS materials in a space environment and has established a program to develop HTS technology for space applications.

An Overview of HTS Applications

The aerospace applications of HTS that are currently being pursued within NASA Lewis Research Center can be grouped into three main thrusts: space communications, power, and propulsion systems. The HTS space communications applications are discussed in another paper in this proceedings. Some of the promising applications, an estimate of the anticipated benefits, and comments on the technology status are discussed below:

Efficient HTS Low-Mass Power Transmission Line: Many future space applications will require large amounts of power to be transmitted over large distances. One example is a nuclear electric space power plant to provide power to a manned space facility. In order to avoid the large amount of mass which would be required to provide radiation shielding if the reactor were located near the facility, the reactor would be some distance from the facility and the power would be transmitted by means of a power transmission line. A study by Hull and Myers[2] has shown that a passively cooled HTS transmission line is feasible using existing technology. For a system 1 km long, the mass of an HTS power transmission system carrying a current density of 5000 A/cm^2 would be approximately one eighth the mass of an aluminum system carrying anywhere between 250 and 1000 A/cm^2. Bulk HTS superconductors with J_c exceeding this valve have already been demonstrated.

Superconducting Magnetic Energy Storage (SMES): The energy storage system of a space vehicle is one of the most massive and voluminous systems on board. For orbital missions the power system must provide power during the eclipse portion of an orbit and serve a load leveling function for optimum use of the power source for both solar and nonsolar based space power systems. Studies have shown that HTS based SMES systems can be competitive, on an energy per unit mass basis, with advanced batteries and fuel cells. But SMES has impressive system advantages over other energy storage technologies in the areas of efficiency, cycle life, and flexibility.[3] These benefits may become significant in areas which are not apparent in the first consideration. For example, a very efficient energy storage system on a low orbit vehicle, such as the proposed space station, could allow a decreased mass and size of the photovoltaic arrays. The decreased size of the arrays would cause less drag and orbital decay on the station, which in turn would mean decreased need for reboosting. This could mean considerable savings over the lifetime of the station.

Determination of hepatic iron in liver by superconducting magnetogradiometer: At the present time the level of iron in the liver can be determined only by biopsy. Lewis Research Center, along with Metro General Hospital of Cleveland, is investigating using a first-order magnetogradiometer made of HTS wire connected by HTS wire to a low-temperature SQUID amplifier to perform a noninvasive measurement of liver iron levels. The HTS wire is required to carry a persistent current of only 1 μA. Other requirements are that the wire be capable of being formed into a coil and be superconducting in the presence of a moderate magnetic field.

Active Superconducting Magnetic Bearings: Lewis Research Center is also performing a study to estimate the performance and losses in active superconducting magnetic bearings and to determine the required properties of the HTS material used to fabricate them. Configurations being studied have both the stator and rotor superconducting, and also the stator superconducting and the rotor is a nonsuperconducting soft magnetic material.

Concluding Remarks

HTS materials can offer many improvements and advantages over conventional technology for both space and ground based applications. A complete system analysis must be done to determine the total benefits from using HTS materials. Although many applications of HTS materials are near term, much materials research remains to be done to improve the properties and determine the reliability of these materials.

Acknowledgments

The authors thank P. Aron, G. Brown, K. Faymon, and S. Levine for their contributions to this paper.

References

1. M. Mirakami, Materials Research Society Spring Meeting, San Francisco, Calif., April 16-21, 1990.

2. J. R. Hull and I. T. Myers, "High Speed Superconductors for Space Power Transmission Lines," Proc. ASME Winter Annual Meeting Conference, San Francisco, Calif., December 10-15, 1989.

3. K. Faymon and S. J. Rudnick, "High Temperature Superconducting Magnetic Energy Storage for Future NASA Missions." Presented at IECEC, 1988.

N91 24081

UNCLAS

N91-24081

SUPERCONDUCTING MICROWAVE ELECTRONICS AT LEWIS RESEARCH CENTER

Joseph D. Warner, Kul B. Bhasin, and Regis F. Leonard
National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135

ABSTRACT

Over the last three years, NASA Lewis Research Center has investigated the application of newly discovered high temperature superconductors to microwave electronics. Using thin films of $YBa_2Cu_3O_{7-\delta}$ and $Tl_2Ca_2Ba_2Cu_3O_x$ deposited on a variety of substrates, including strontium titanate, lanthanum gallate, lanthanum aluminate and magnesium oxide, a number of microwave circuits have been fabricated and evaluated. These include a cavity resonator at 60 GHz, microstrip resonators at 35 GHz, a superconducting antenna array at 35 GHz, a dielectric resonator filter at 9 GHz, and a microstrip filter at 5 GHz. Performance of some of these circuits as well as suggestions for other applications are reported.

INTRODUCTION

Investigations to determine space electronics applications of high temperature superconductor (HTS) at NASA Lewis Research Center were initiated soon after the discovery of superconductivity in ceramic oxides $La_{1-x}Sr_xCu\,O_y$[1] and $YBa_2Cu_3O_{7-\delta}$[2] with transition temperatures 35 K and 93 K. Soon after their discoveries were made others found the Bi-Sr-Ca-Cu-O[3] and Tl-Ca-Ba-Cu-O[4] class of 100 K + superconductors. Since that time the properties of these superconductors became known and we concentrated our efforts on development of thin films of $YBa_2Cu_3O_{7-\delta}$ and $Tl_2Ca_2Ba_2Cu_3O_x$ for space microwave applications. The use of HTS films in a microwave system requires development of thin films on microwave substrates which then can be patterned into desired microwave circuits like filters, phase shifters, ring resonators and delay lines. Such circuits are used in space communication, radar, and sensing systems.[5] Small size, low loss, low power and light weight are desirable features for these circuits.

In this paper, we describe the development of high quality $YBa_2Cu_3O_{7-\delta}$ (YBCO) and $Tl_2Ca_2Ba_2Cu_3O_x$ (TCBCO) thin films on microwave substrates carried out at NASA Lewis and also at various sponsored and cooperative facilities. The method of fabricating and the evaluation of various microwave passive circuits is presented. At the end future applications are also highlighted.

DEVELOPMENT OF THIN SUPERCONDUCTING FILMS

To obtain high quality superconducting thin films on suitable substrates for microwave applications the substrate lattice constants must be closely matched to those of the films and there must not be a detrimental chemical reaction between the substrates and the films. In addition, the film composition must be as close to the correct composition as possible. To date, very high quality films have been obtained by using several physical and chemical deposition techniques. Many of these techniques had required postannealing at high temperatures. This high temperature anneal causes chemical interactions at the film-substrate interface, making the film-substrate interface unsuitable for microwave application.[6] To circumvent this problem, a laser ablation technique[7,8] followed by an in situ annealing procedure has been pursued and developed for the growth of YBCO films. Also, for the growth of TCBCO films, sputtering and laser ablation techniques were chosen and funded at University of Cincinnati and University of Nebraska at Lincoln, respectively. Though, with both methods an ex-situ anneal with thallium over pressure was required.

Both YBCO and TCBCO films were evaluated for their microwave properties. The reasons why YBCO was selected were because it can be grown single phase easily and has no competing phases when the oxygen pressure is 1 atm or less. Its critical field can exceed 100 T at low temperatures and its critical current is greater than $2x10^6$ A/cm^2 at 77 K. TCBCO films were chosen because the TCBCO phase has a transition of 125 K in bulk form and early properties of bulk TCBCO showed that the material had lower 1/f noise than YBCO. The progress to date on the growth and characterisation of YBCO and TCBCO thin films is presented subsequently.

$YBa_2Cu_3O_{7-\delta}$ Films

At the beginning of the research three methods of producing YBCO films were investigated. They were sequential evaporation, coevaporation, and laser ablation. In the following paragraphs, we briefly describe the techniques and give references for them.

Produced by sequential evaporation. - One of the earliest attempts at The Ohio State University to produce YBCO films was by sequential evaporation of copper, barium fluoride, and yttrium and followed by a postanneal.[9] This procedure did produce YBCO films that were mainly a-axis aligned, but the films generally were poor having low transition temperatures (T_c) and critical current density (J_c) values and high porosity.[10] Nevertheless, this technique did allow us to set up the necessary equipment for investigation of the optical and microwave properties of HTS films.[9,11,12]

Produced by coevaporation. - Coevaporation is being pursued jointly by Oberlin College and NASA Lewis. The film by this technique is grown at room temperature by coevaporating Cu, BaF_2, and Y from separate electron beam guns.[13] Then the film is postannealed similar to the postannealed sequentially evaporated films. One of the advantages for coevaporation is that patterns with 2 μm lines can be formed by using photoresist and a lift-off technique. So far the films produced have had a high T_c of 90 K but J_c in the $5x10^5$ A/cm^2 range at 77 K. The low J_c is due to the high temperature annealing conditions which cause a substantial amount of a-axis growth and porosity.

Produced by laser ablation. - The laser ablation technique[7] has given us the best YBCO films to date. Films by this technique were produced at NASA Lewis and the technique has also been set up at The Ohio State University, successfully. The basic principle of laser ablation is that a short wavelength and short pulse duration laser beam is focused onto a YBCO target. This evaporates the surface of the target and produces a plasma plume. Since the pulse duration is very short there is very little heat transferred to the target preventing thermal melting of the target and the noncongruent evaporation of the individual atoms. Therefore, a stoichiometric composition of the target is ablated from the surface.

The best films on $LaAlO_3$ were c-axis aligned and had a T_c around 90.6 K immediately after deposition as determined by a standard four point resistance measurement. Resistance versus temperature behavior for a YBCO film is shown in fig. 1. Critical current density J_c versus temperature is shown in fig. 2. As can be seen, the value of J_c was $2x10^6$ A/cm^2 at 77 K. The surface of the films was very smooth with some small structure of about 0.25 μm in size. This size of structure has been confirmed by scanning tunneling microscopy. In table I, we list the performance of YBCO and TCBCO thin films on various microwave substrates along with the microwave properties of these substrates.

$Tl_2Ca_2Ba_2Cu_3O_x$ Films

TCBCO films on $LaAlO_3$ substrates have been made at University of Cincinnati and University of Nebraska at Lincoln. University of Cincinnati made their films by r.f. sputtering of a compressed 2223 powder target[14,15] followed by an anneal in Tl vapors. The smooth, dense films have a T_c between 103 K to 107 K with a 6 K transition width. The films are a mixture of 2223 and 2212 phases. University of Nebraska made their films by laser ablation and have achieved a T_c as high as 115 K, see fig. 3, but typical T_c were around 108 K.[16] But the films are not smooth and have some porosity. They used a compressed powder target of TlO_2, $BaCuO_2$, and $CaCuO_2$. Presently, they are exploring the use of a 2223 target to achieve a film that is denser and has a T_c above 120 K.

Microwave Characterisation

Surface resistance (R_s) of superconducting film is a basic physical property that can be used to determine the quality of the film and is necessary for microwave device design. Currently, surface resistance values are obtained by cavity[17,18] and stripline measurements[19]. These measurements are time consuming, and it would be worthwhile to correlate R_s with dc or low frequency measurements, such as dc conductivity above T_c, magnetic penetration depth, T_c, and a.c. susceptibility measurement; but so far there has been no consistent correlation reported between any of these measurements and R_s. The microwave conductivity (σ) is another physical property that can be measured. The conductivity for a superconductor is complex ($\sigma = \sigma_1 + i\sigma_2$) when the temperature is below T_c and it can be related to R_s and to the penetration depth (λ)[20].

Miranda et al.[12] have determined σ by measuring the microwave power transmitted through superconducting thin films in a waveguide experiment. From the value of σ_2 the magnetic penetration depth (λ) can be obtained. A summary of results of σ_1, λ_o, R_s for YBCO film on various substrates is shown in table II. The R_s for these films were calculated from σ_1, and λ. For YBCO on $LaAlO_3$ the R_s was $1.4 \times 10^{-3}\ \Omega$ which compares very well with other data.[21] The surface resistance is an order of magnitude lower than that of copper up to 60 GHz.[21] This demonstrates that with microwave transmission measurements one can obtain the R_s, λ, and σ of HTS films. These three properties are necessary for the determination of the quality of HTS films and the designing of microwave circuits.

FABRICATION

We have developed a method to fabricate microwave circuits and test devices for HTS films. The method is to use standard photolithography using negative photoresist and a "wet" chemical etchant. This etchant was either a 1 at % solution of bromine in ethanol or dilute phosphoric acid in water. In addition, circuits identical to the HTS circuits were made from patterned gold. This allows a direct comparison between HTS and gold circuits.

APPLICATION INVESTIGATIONS

The application of HTS to communication or radar systems does not only depend on the properties of HTS material but also the total system cost versus performance. The single most important issue for space communication is the possible need for cryogenic cooling. If HTS circuits must be cooled cryogenically then that cost in terms of power and weight must be considered along with the improved performance of the devices and any savings in weight and power from the replacement of standard equipment with HTS equipment. However, there are some missions, such as deep space probes that only would require passive cooling or other missions where cryogenic cooling is necessary for other functions, where this would not be an issue. Some of the microwave applications and testing are discussed below. A brief description of each is made or referred to. The applications under investigation are resonators, filters, phase shifters, hybrid circuits, and antennae.

60 GHz Cavity

The surface resistance of HTS films can be calculated directly from the "Q" value of a resonate cavity experiment where one of the end walls of the cavity is replaced with the HTS film.[18] This has been done at 60 GHz for YBCO films on $LaGaO_3$ and $SrTiO_3$. The R_s values for the two films can be seen in fig. 5. These films were approximately 1 μm thick; therefore, being polycrystalline c-axis oriented films. The results obtained agreed with other[21] results for YBCO films grown by laser ablation and sputtering but measured at different frequencies. This agreement confirms the f^2 dependence of R_s and that the microwave properties for polycrystalline films of YBCO is independent of whether the films are grown on $SrTiO_3$ or $LaGaO_3$.

Resonator Circuits

HTS stripline ring resonators were fabricated on $LaAlO_3$ substrates.[22] For measurement the resonators were mounted in a cosine tapered ridge waveguide to microstrip test fixture as shown in fig. 4. That structure was cooled by a closed cycle refrigerator. The resonators were measured by an HP8510 automatic network analyser. The resonance frequency of the resonator changed rapidly with temperature just below T_c. This change was due to the change in the circuit reactance caused by the change in the magnetic penetration depth.[22]

The best resonators measured to date and compared to gold is shown in fig. 6. The unloaded "Q" ranged from 2500 to 1000 at 20 and 77 K, respectively. This corresponds to a surface resistance value of 8 mΩ at 77 K at 35 GHz, a value two to three times better than copper at the same temperature and frequency.

Filters

Another candidate for application of HTS thin films is in the area of passive microwave filters. Taking advantage of the low losses for HTS film we have considered where they could be applied within a satellite transponder to improve performance. Based on results obtained to date on the performance of superconducting microstrip resonator circuits with high "Q" values as compared to an all metal microstrip resonators, we project the application of superconducting passive circuits as low loss, high "Q" filters[13], high "Q" resonators, delay lines, power splitter, power combiners, and resonator stabilized oscillators.

Phase Shifters

In addition to these applications, extremely low loss phase shifters using superconducting switches are also feasible. In fig. 7, we show a phase shifter which utilizes superconducting-normal-superconducting switches in place of FET/diode switches. The switches are fabricated from high temperature thin films of YBCO. The switches operate in the bolometric mode with the film near its transition temperature. Radiation from a light source raises the temperature higher than the film's T_c and consequently causes the film to become resistive. When the light is "on" the microwave signal travels past the switch, but is reflected when the light is off. To achieve the desired phase shift, the pair switches on the same side is illuminated. Fig. 8 shows the predicted behavior for a 180° phase shifter with a R_s value that is the same as gold at 77 K and having a R_s of 100 $\mu\Omega$ in the normal state. It has an with exceptional narrow insertion loss envelope and excellent return loss.

Hybrid Semiconductor/Superconductor Device

The natural use of hybrid semiconductor and superconductor devices will be where III-V compound semiconductors, such as AlGaAs, InGaAs, and III-V hetrojunction material, will be used at temperature around 77 K for devices that cannot operate or give the necessary performance at room temperatures. Some of these applications are for low noise amplifies at frequencies above 22 GHz or solid state amplifiers above 70 GHz. Since these semiconducting devices will have to be cooled to 77 K there is no penalty in terms of the cost or reliability of the refrigeration to be paid to use superconductor devices. Therefore, it is natural to use HTS circuits in conjunction with III-V semiconductors devices to obtain the best performance of devices at these temperatures.

In fig. 9, we show an example of hybrid semiconductor/superconductor device for an ultra low noise receiver for satellite applications. This receiver take the advantage of the excellent noise properties of AlGaAs HEMT technology and the low noise and resonator properties of superconducting transmission lines to achieve the ultra low noise and stable amplification.

Superconducting Phased Arrays

Superconducting antennas have long been imagined as extremely low loss devices. The use of superconductors to reduce the size of antennas to a fraction of a wavelength and to make super-directive arrays have been some of the more popular subjects. It has been show recently, however, that of these various uses, the most practical use of superconductors in antennas will be in microwave and millimeter wave antennas.[23] To demonstrate the use of superconductors in such antennas, current research at NASA Lewis, in cooperation with Ball Aerospace, is focusing upon the fabrication and testing of a four element planar array at 30 GHz (fig. 10). The performance of this antenna will be compared to an identical array fabricated using gold instead of HTS.

CONCLUSIONS

We have demonstrated that rare-Earth oxide thin superconducting films can be deposited on various microwave substrates with critical temperature T_c above 77 K, critical current densities J_c above 10^6 A/cm^2, and low surface resistance. The films can be easily etched into microwave transmission line circuits. Microwave circuit ring resonator fabricated from a YBCO superconducting film on $LaAlO_3$ substrate showed "Q" values four time that for similar resonator made from a gold film. Several key HTS circuits such as filters, oscillators, phase shifters, and phased array antenna feeds are feasible in the near future. For technology to improve further, reproducible, large area films have to be grown on low dielectric constant, low loss microwave substrates. Tradeoffs between superconducting microwave circuits with cryogenic systems and normal metal microwave circuits will have to be quantitatively established to determine their suitability for advanced communication and sensor systems.

REFERENCES

1. Bednorz, J.G.; and Muller, K.A.: Possible High T_c Superconductivity in the Ba-La-Cu-O System. Z. Phys. B. Condensed Matter, vol. 64, 1986, pp. 189-193.

2. Chu, C.W., et al.: Evidence for Superconductivity Above 40 K in the La-Ba-Cu-O Compound System. Phys. Rev. Lett., vol. 58, no. 4, Jan. 26, 1987, pp. 405-407.

3. Maeda, M., et al.: A New High-T_c Oxide Superconductor Without a Rare Earth Element. Jpn. J. Appl. Phys. Lett., vol. 27, no. 2, Feb. 1988, pp. L209-L210.

4. Sheng, Z.Z.; and Herman, A.M.: Bulk Superconductivity at 120 K in the Ti-Ca/Ba-Cu-O System. Nature, vol. 332, no. 6160, Mar. 10, 1988, pp. 138-139.

5. Heinen, V.O., ed.: Superconductivity Applications for Infrared and Microwave Devices. Proc. SPIE Vol. 1292, SPIE, Bellingham, WA, 1990.

6. Bhasin, K.B., et al.: High-Temperature Superconducting Thin Film Microwave Circuits: Fabrication, Characterisation, and Applications. Superconductivity Applications for Infrared and Microwave Devices, Proc. SPIE Vol. 1292, SPIE, Bellingham, WA, 1990 (Also, NASA TM-103235).

7. Warner, J.D.; Meola, J.E.; and Jenkins, K.A.: Study of Deposition of $YBa_2Cu_3O_{7-x}$ on Cubic Zirconia. NASA TM-102350, 1989.

8. Bhasin, K.B., et al.: Performance and Modeling of Superconducting Ring Resonators at Millimeter-Wave Frequencies. 1990 IEEE MTT-S International Microwave Symposium Digest, Vol. 1, IEEE, 1990, pp. 269-272.

9. Valco, G.J., et al.: Photoresponse of $YBa_2Cu_3O_{7-\delta}$ Granular and Epitaxial Superconducting Thin Films. NASA TM-103144, 1990.

10. Stan, M.A., et al.: Transport Measurements on Granular Y-Ba-Cu-O Films. Physica C, vol. 162-164, 1989, pp. 365-366.

11. Valco, G.J., et al.: Sequentially Evaporated Thin Y-Ba-Cu-O Superconducting Films on Microwave Substrates. NASA TM-102068, 1989.

12. Rohrer, N.J., et al.: Sequentially Evaporated Thin Film $YBa_2Cu_3O_{7-x}$ Superconducting Microwave Ring Resonator. NASA TM-103180, 1990.

13. Mankiewich, P.M., et al.: Reproducible Technique for Fabrication of Thin Films of High Transition Temperature Superconductors. Appl. Phys. Lett., vol. 51, no. 21, Nov. 23, 1987, pp. 1753-1755.

14. Subramanyam, G., et al.: Fabrication and Chemical Composition of RF-Magnetron Sputtered Ti-Ca-Ba-Cu-O High T_c Superconducting Thin Films. J. Appl. Phys., vol. 68, no. 3, 1990, pp. 1157-1163.

15. Subramanyam, G.; Radpour, F.; and Kapoor, V.J.: Fabrication of Ti-Ca-Ba-Cu-O Superconducting Thin Films on $LaAlO_3$ Substrates. Appl. Phys. Lett., vol. 56, no. 18, Apr. 30, 1990, pp. 1799-1801.

16. Eri[illegible]gton, K.B.; and Ianno, N.J.: Thin Films of Uniform Thickness by Pulsed Laser Deposition. To be published in Journal of Applied Physics.

17. Miranda, F.A., et al.: Microwave Conductivity of Superconducting Bi-Sr-Ca-Cu-O Thin Films in the 26.5 to 40.0 GHz Frequency Range. Physica C, vol. 168, nos. 1-2, June 1, 1990, pp. 91-98.

18. Miranda, F.A., et al.: Millimeter-Wave Surface Resistance of Laser Ablated $YBa_2Cu_3O_{7-\delta}$ Superconducting Films. Appl. Phys. Lett., vol. 57, no. 10, Sept. 3, 1990, pp. 1058-1060.

19. Bhasin, K.B., et al.: Determination of Surface Resistance and Magnetic Penetration Depth of Superconducting $YBa_2Cu_3O_{7-\delta}$ Thin Films by Microwave Power Transmission Measurements. NASA TM-103616, 1990 (to be published in, IEEE Trans. Magnetics, 1991).

20. Bhasin, K.B., et al.: Microwave Conductivity of Laser Ablated $YBa_2Cu_3O_{7-\delta}$ Superconducting Films and its Relation to Mirostrip Transmission Line Performance. AMSAHTS 1990: Advances in Materials Science and Applications of High Temperature Superconductors, Y. Flom, ed., NASA CP-10043, 1990, pp. 78-81.

21. Klein, N., et al.: Millimeter Wave Surface Resistance and London Penetration Depth of Epitaxially Grown $YBa_2Cu_3O_{7-x}$ Thin Films. Physica C, vols. 162-164, pt. II, 1989, pp. 1549-1550.

22. Miranda, F.A., et al.: Complex Permittivity of Lanthanum Aluminate in the 20 to 300 K Temperature Range from 26.5 to 40.0 GHz. Microwave Opt. Technol. Lett., vol. 3, no. 1, Jan. 1990, pp. 11-13.

23. Hansen, R.C.: Superconducting Antennas. IEEE Trans. Aerospace Electron., vol. 26, no. 2, Mar. 1990, pp. 345-355.

TABLE I. - KEY PROPERTIES OF MICROWAVE SUBSTRATE MATERIALS AND THE TRANSITION TEMPERATURE (T_c) FOR LASER ABLATED $YBa_2Cu_3O_7$ FILM ON THE VARIOUS SUBSTRATES

Material	Laser, T_c (K)	Dielectric constant	Loss tangent	Lattice size, Å
Magnesium oxide (MgO)	88	9.65	$4x10^{-4}$	4.178 (100)
Lanthanum aluminate ($LaAlO_3$)	90	22	$5.8x10^{-4}$	3.792 (110)
Lanthanum gallate ($LaGaO_3$)	88	27	$2x10^{-3}$	3.892 (110)
Sapphire (Al_2O_3)	73	9.4	$1x10^{-6}$	5.111 (011)
Yttria stabilized zirconia (ZrO)	89	11.6 27	$6x10^{-4}$	3.8795 (100)
Silicon (Si)	--	12	$10x10^{-4}$	5.43 (100)
Gallium arsenide (GaAs)	--	13	$6x10^{-4}$	5.563 (100)

TABLE II. - THE REAL PART OF THE CONDUCTIVITY $\sigma = \sigma_1 + i\sigma_2$, AND AT 77 K AND 33 GHz, THE CALCULATED R_s AT 5 GHz, THE MAGNETIC PENETRATION DEPTH (λ_o) FOR $YBa_2Cu_3O_7$ FILMS ON $LaAlO_3$, MgO, AND YTTERIA STABILIZED CUBIC ZIRCONIA (YSZ)

Substrate	Thickness, nm	σ_1, s/m	R_s (Ω), 33 GHz	R_s (Ω), 5 GHz	λ_o, nm
$LaAlO_3$	177	$2.5x10^5$	$1.8x10^{-3}$	$41x10^{-6}$	360
MgO	350	$1.2x10^5$	$5.9x10^{-3}$	$135x10^{-6}$	530
YSZ	120	$2.4x10^5$	$2.5x10^{-3}$	$57x10^{-6}$	590

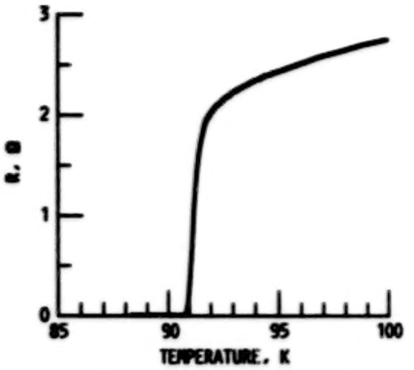


FIGURE 1. - D.C. RESISTANCE VERSUS TEMPERATURE OF LASER ABLATED $YBa_2Cu_3O_{7-\delta}$ FILM ON $LaAlO_3$.

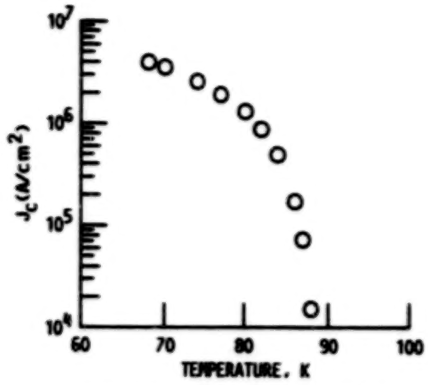


FIGURE 2. - CRITICAL CURRENT DENSITY OF LASER ABLATED $YBa_2Cu_3O_{7-\delta}$ FILM ON $LaAlO_3$.

ORIGINAL PAGE
BLACK AND WHITE PHOTOGRAPH

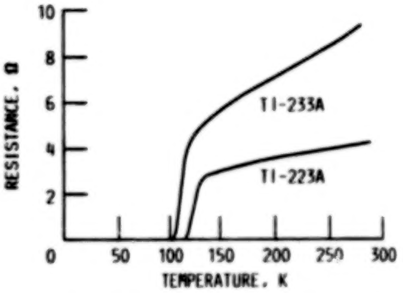


FIGURE 3. - D.C. RESISTIVITY VERSUS TEMPERATURE OF TWO $Tl_2Ca_2Ba_2CuO_x$ FILMS.

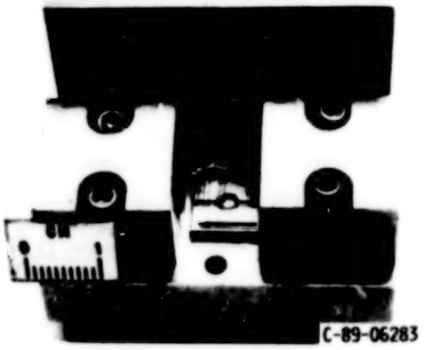

C-89-06283

FIGURE 4. - WAVEGUIDE TEST FIXTURE USED FOR THE MEASUREMENT OF "Q" VALUES OF SUPERCONDUCTING RING RESONATORS.

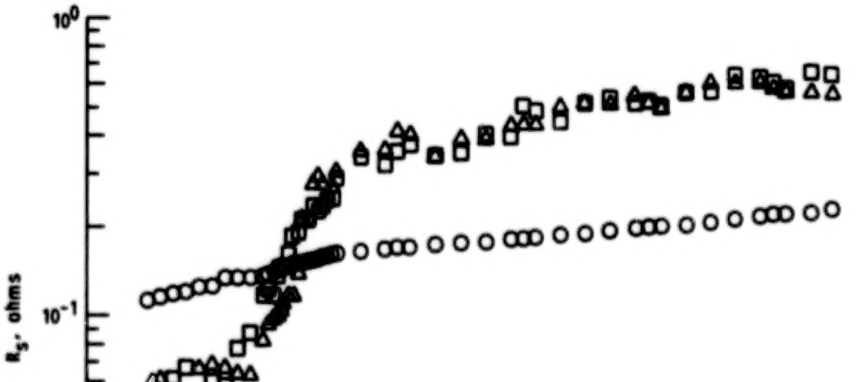


FIGURE 5. - SURFACE RESISTANCE (R_s) VERSUS TEMPERATURE FOR GOLD AND LASER ABLATED $YBa_2Cu_3O_{7-\delta}$ FILMS ON THE INDICATED SUBSTRATES AS MEASURED IN THE 58.6 GHz GOLD CAVITY.

ORIGINAL PAGE IS
OF POOR QUALITY

ORIGINAL PAGE IS OF POOR QUALITY

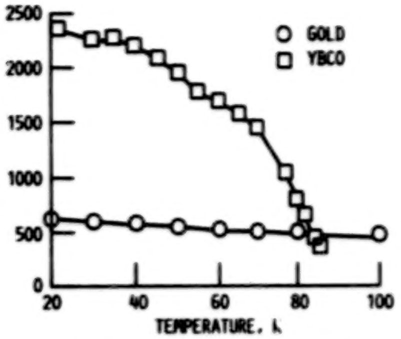


FIGURE 6. - 35 GHz RING RESONATOR UNLOAD "Q" VERSUS TEMPERATURE FOR GOLD AND LASER ABLATED $YBa_2Cu_3O_{7-\delta}$ FILM ON $LaAlO_3$.

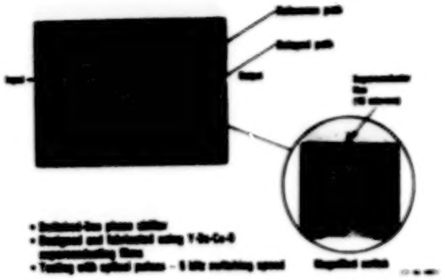
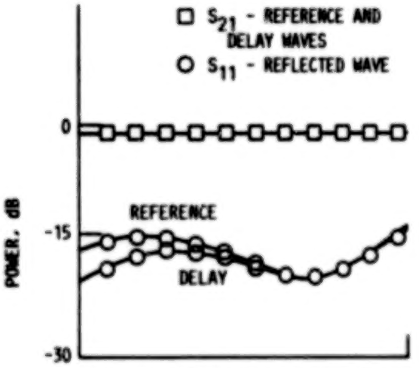
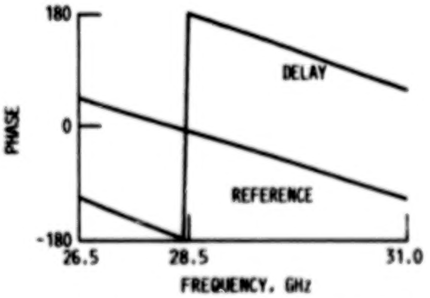


FIGURE 7. - OPTICALLY CONTROLLED SUPERCONDUCTOR PHASE SHIFTER.



(a) S_{21} AND S_{11} FOR REFERENCE, DELAY, AND REFLECTED WAVES.



(b) PHASE RESPONSE VERSUS FREQUENCY FOR REFERENCE AND DELAY PATHS.

FIGURE 8. - THEORETICAL RESPONSE FOR THE DESIGNED OPTICALLY CONTROLLED $YBa_2Cu_3O_{7-\delta}$ PHASE SHIFTER AT 77K.

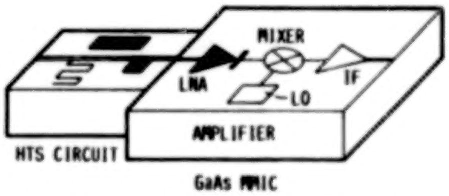
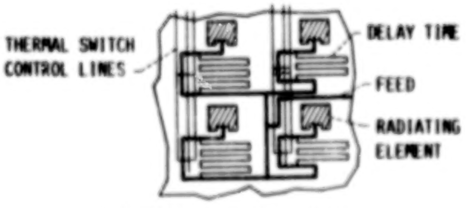

MIXER
LNA
IF
LO
AMPLIFIER
HTS CIRCUIT
GaAs MMIC

FIGURE 9. - SUPERCONDUCTING GaAs MMIC HYBRID RECEIVER.


THERMAL SWITCH CONTROL LINES
DELAY TIME
FEED
RADIATING ELEMENT

4x4 ARRAY PRELIMINARY DESIGN

FIGURE 10. - PRELIMINARY DESIGN OF A THERMALLY SWITCHED PHASE ARRAY SUPERCONDUCTING ANTENNA FOR A 4x4 ARRAY.

SUPERCONDUCTIVE WIRES AND DEVICES FOR CRYOGENIC APPLICATIONS

Dr. John Buckley
Langley Research Center
Hampton, VA 23665

DOCUMENTATION OF THIS PAPER WAS NOT PROVIDED FOR INCLUSION IN THESE PROCEEDINGS. FOR FURTHER INFORMATION, PLEASE DIRECT ALL INQUIRIES TO THE NAME AND ADDRESS LISTED ABOVE.

N91

24082

UNCLAS

N91-24082

MELT-PROCESSED BULK SUPERCONDUCTORS: FABRICATION AND CHARACTERIZATION FOR POWER AND SPACE APPLICATIONS

Hamid Hojaji, Aaron Barkatt and Shouxiang Hu
The Catholic University of American
Washington, D.C. 20059

Arthur N. Thorpe, Matthew F. Ware and David Davis
Howard University
Washington, D.C. 20059

Sidney Alterescu
NASA Goddard Space Flight Center
Greenbelt, MD 20910

ABSTRACT

Melt-process bulk superconducting materials based on variations on the base $YBA_2Cu_3O_x$ have been produced in a variety of shapes and forms. Very high values of both zero-field and high-field magnetization have been observed. These are useful for levitation and power applications. Magnetic measurements show that the effects of field direction and intensity, temperature and time are consistent with an aligned grain structure with multiple pinning sites and with models of thermally activated flux motion.

HIGH-MAGNETIZATION SUPERCONDUCTING CUPRATES; FORMULATION, PROCESSING AND MICROSTRUCTURE

Recent studies conducted at the authors' laboratories have shown that the combination of melt processing of Y-Ba-Cu-O superconducting materials with the use of non-stoichiometric formulations makes it possible to produce materials with very high magnetization [1, 2]. Zero-field dc magnetization values of up to 18 emu/g and remanent magnetization (M_+ - M) values of up to 21 emu/g at a field intensity of 5 kOe have been observed at 77 K. The non-stoichiometric compositions mentioned above involve additions of excess yttrium, rare earths such as Gd, Tb, Ho and Yb, or transition metals such as Nb, to the base $YBa_2Cu_3O_x$ composition. The high magnetization of these materials has been concluded to be due to the controlled formation of a highly oriented grain structure, interspersed with second phase sites (in particular, Y_2BaCuO_5) which promote flux pinning. This approach has been shown suitable for the production of large pieces which have been fabricated to give a variety of shapes, including cylinders, bars, plates,a nd prototype bearing elements. Measurements of levitation forces and of magnetic stiffness gave high values for these materials, which were correlated with their high magnetization [3]. This makes these materials attractive for levitation applications such as levitation bearings. In addition, the approach described above has been shown to be suitable for the controlled production of materials with different types of hysteresis loops depending on the composition of the samples and on the procedure used to prepare them. In particular, it is possible to obtain, on one hand, materials with a high degree of flux trapping, i.e., a large remanent magnetization in zero field, and, on the other hand, materials which effectively pin the flux at higher fields. This increases the scope of applications of the materials to include, for instance, magnetic energy storage.

The melt-processed samples described above were recently subjected to extensive measurements of their magnetic properties, with an emphasis on flux trapping and flux creep, and on their magnetomechanical behavior as related to levitation and suspension.

ANISOTROPIC MAGNETIC FEATURES AND FIELD DEPENDENCE

The measurement of dc magnetization and magnetic susceptibility was described in earlier paper [1, 4]. All the data in the present paper relate to a single melt-processed sample of $YBa_2Cu_3O_x$ without any additional dopants. The magnetization values obtained with this sample are somewhat lower than those

C-5

obtained recently with the doped samples mentioned above; detailed magnetic studies of the latter samples are currently in progress.

It has been recently found that even though the melt-processed samples are polycrystalline, the effect of their grain orientation with respect to the applied magnetic field is very important. The effect of varying the orientation of a high-magnetization sample on the magnetization values at 77 K is shown in Figure !. It can be seen that rotation of the sample at an angle of 90^0 with respect to the direction at which the maximum magnetization was observed caused the magnetization to drop by a factor of approximately 3.5, viz. from 15 emu/g to 4.3 emu/g. The combination of results obtained at the two extreme orientations with those obtained at other angles gives excellent agreement with a simple two-component model, according to which

$$M_r = M_r \| + (M_r \perp - M_r \|) * \cos^2 \theta \tag{1}$$

where M_r is the magnetization measured when the applied magnetic field, oriented at an angle Θ with respect to the c axis of the $YBa_2Cu_3O_7$ grains, is removed. $M_r \|$ and $M_r \perp$ are the zero-field magnetization values obtained after applying fields parallel and perpendicular to the c axis or ab plane, respectively. The agreement between the model and the observed results is shown in Figure 2. The results show an excellent linear correlation between M_r and $\cos^2 \Theta$. The angular dependence of the magnetization observed in the present case is consistent with the behavior of the flux density in a layered superconductor with weak coupling between adjacent layers as analyzed by Kes et al [5]. However, in the present case better linear correlation was observed between M_r and $\cos^2 \Theta$ than between M_r and $\cos \Theta$.

The dependence of the remanent magnetization at 77 K on the maximum applied magnetic field is shown in Figure 3 for field intensities of up to 10 kOe. Both the values obtained in the direction of the c axis and those measured in the direction perpendicular to it are included. It can be seen that the magnetization observed in both directions levels off at fields of 2-3 kOe. A measurement of the initial slop of the M - H dependence was performed using very small applied fields. Under these conditions, the relationship between M and H is linear, as shown in Figure 4, and the values of the residual magnetization after removal of the field are very low, as shown in Figure 5.

ANISOTROPY IN THE CRITICAL CURRENT DENSITIES

Bean's model [6] has been extensively used to obtain critical current densities, J_c, magnetization curves. Good agreement has been observed in several studies between the observed transport current density and the predictions of Bean's model [5a, 5b]. Recently, Hu et al [5c], developed an extended form of Bean's model for the specific case of a grain-aligned bulk superconductor. A simplified version of the equation in the case where the applied field H is parallel to the c-axis is given by:

$$\Delta M = \frac{1}{30} J_c^{ab} d \tag{2}$$

where J_c^{ab} is the critical current density parallel to the ab plane and d is the smaller dimension of the grains in the ab plane. With d = 0.2 cm, Δ M = 190 emu/cm^3 at H = 34 Oe, and T = 77 K, J_c^{ab} is calculated to be equal to 2.8 X 10^4 A/cm^2. At H = 2 kOe, J_c^{ab} is equal to 1.6 X 104 A.cm^2.

When the applied field is perpendicular to the c-axis, the calculated J values decrease to 7.3 X 10^3 and 2.9 X 10^3 A/cm^2 at 300 Oe and a 2 kOe, respectively.

The anisotropy in J_c, i.e., the ratio between the J_c values corresponding to the two directions of the applied field, is about 4 at low fields, and it increases to about 5.5 at 2 kOe. Compared with the anisotropy of the single crystal [5d], which is about 175 at 40 K and H = 8 kOe, the much lower degree of anisotropy of the melt processed material characterized in this work is attributed to the much larger number of pinning sites uniformly distributed throughout the material.

DEPENDENCE OF THE MAGNETIZATION ON TEMPERATURE AND TIME

Upon the dependence of the of the dc magnetic susceptibility and magnetization on temperature, it is observed that the values of these two magnetic properties decrease as the temperature is raised and approaches T_c. The data in Figure 6 represent magnetic susceptibility measurements in a very low field (110 Oe) at increasingly higher temperatures following the application and removal of a 2.5-kOe field at the lowest temperature of 77 K. Figure 7 likewise represents magnetization values obtained upon cooling to 66 K, applying a 10-kOe field and removing it, and then gradually raising the temperature of the sample. It is important to note that when the warm-up is interrupted and the temperature is lowered before being raised again, the magnetic susceptibility and the remanent magnetization remain constant until the temperature increases beyond the point at which the warm-up was interrupted. This shows that these experiments give a true measure of the dependence of the capability of the samples to trap magnetic flux, and that the results are not significantly affected by the interruptions in the heating schedule. The ensuing data can be interpreted as reflecting the role of thermal activation in flux trapping. (It should be noted that the apparent leveling off of the values of the magnetic properties at the lowest temperatures employed in the measurements shown in Figures 6 and 7 is an experimental artifact).

The temperature effects on the apparent activation energy of flux trapping are not strongly dependant on the orientation. Upon measuring the dc magnetization in the direction which yields the highest values, i.e., the direction perpendicular to the ab planes, the zero-field readings on the sample which was used to obtain the data in Figure 7 at temperatures of 77 K and 67 K were 15.3 and 37.6 emu/g, respectively. The corresponding readings in the direction parallel to the ab planes were 7.2 emu/g and 12.5 emu/g. The resulting ratio between $M_r \perp$ and $M_r \parallel$ was 2.99 at 77 K and 3.02 at 67 K. The observed temperature dependence in the direction perpendicular to the ab planes between 66 and 79 K can be described in terms of an activation energy of 3.86 kJ/mol or 0.0399 eV. A very similar activation energy was obtained when the temperature dependence was obtained when the residual magnetization at an applied field of 2.4 kOe was measured instead of the zero-field magnetization. It is important to note that the near-linear dependence of the magnetization on the temperature breaks down as the critical temperature T_c is approached and is replaced by a tailing curve with progressively decreasing slope. Similar behavior has been observed upon examining the temperature dependence of the critical current density J_c in thin films of $YBa_2Cu_{3O}x$ and flux creep [12] and which has been shown to be relatively insensitive to the presence of weak links [11].

Long-term measurements of the residual zero-field magnetization at a constant temperature of 77 K were carried out in order to characterize the rate of its decay as a result of flux creep. The results are shown in Figure 8. It can be seen that the decay of the magnetization in each of the two directions specified above follows a logarithmic rate law of the type:

$$M_r(t) = M_r(0)*[1-kT/E*\ln(1+t/t_0)] \qquad (3)$$

where $M_r(o)$ and $M_r(t)$ are the magnetization values observed initially and after time t, respectively, E is the activation energy obtained above, and t_0, is a characteristic decay time, amounting approximately $2*10^3$ seconds at 77 K. This model was originally developed by Hagen et al [13, 14] to describe thermally activated flux motion in high-T_c superconductors consisting of many pinning regions.

In conclusion, the decay of M_r with time in high-magnetization materials produced by melt processing of on-stoichiometric variants of $YBa_2Cu_3O_7$, together with the dependence of the magnetization on temperature and on grain orientation, are all consistent with the picture of a structure with multiple pinning regions which is subject to thermally activated flux motion. Furthermore, this picture is consistent with the oriented grain structure of these materials [1, 2], which contains micron- and submicron-sized second phase regions likely to provide sites for flux pinning.

ACKNOWLEDGEMENTS

This study was supported by the National Aeronautics and Space Administration under Contract No. NAG 5-1017.

REFERENCES

1. Hojaji, H., Barkatt, A., Michael, K. A., Hu. S., Thorpe, A. N., Ware, M. F., Talmy, I. G., Haught, D. A., and Alterescu, S.: Yttrium Enrichment and Improved Magnetic Properties in Partially Melted Ya-Ba-Cu-O Materials. J. Mater. Res. Vol. 5, No. 4, 1990, pp. 721-730.

2. Hojaji, H., Barkatt, A., Hu, S., Michael, K. A., Thorpe, A. N., Talmy, I. G., Haught, D. A., and Alterescu, S.: Superconducting Cuprates Prepared by the Melt Quench Process and Containing Excess Y or Additives. Mat. Res. Bull. Vol. 25, No. 6, 1990, pp. 765-777.

3. Moon, F. C., Chang, P. -Z., Hojaji, H., Barkatt, A., and Thorpe, A. N.: Levitation Forces, Relaxation and Magnetic Stiffness of Melt-Quenches $YBa_2Cu_3O_x$. Japan. J. Appl. Phys., in press.

4. Hojaji, H., Michael, K. A., Barkatt, A., Thorpe, A. N., Ware, M. F., Talmy, I. G., Haught, D. A., and Alterescu, S.: A Comparative Study of Sintered and Melt-Grown Recrystallized $YBa_2Cu_3O_x$. J. Mater. Res. Vol. 4, No. 1, 1990, pp. 28-32.

5. Kes, P. H., Aarts, J., Vinokur, V. M., and van der Beek, C. J.: Dissipation in Highly Anisotropic Superconductors. Phys. Rev. Lett. Vol. 64, No. 9, 1990, pp. 1063-1066.

6. Bean, C. P.: Magnetization of Hard Superconductors. Phys. Rev. Lett. Vol. 8, No. 6, 1962, pp. 250-253.

7. Murakami, M., Morita, and Koyama, N.: Magnetization of a $YBa_2Cu_3O_7$ Crystal Prepared by the Quench and Melt Growth Process. Japan. J. Appl. Phys. Vol. 28, No. 7, 1989, pp. L1125-L1127.

8. Salama, K, Selvamanickan, V., Gao, L., and Sun, K.: High Current Density in Bulk $YBa_2Cu_3O_x$ Superconductor. Appl. Phys. Lett. Vol. 54, No. 23, 1989, pp. 2352-2357.

9. Hu, S., Hojaji, H., Barkatt, A., Boroomand, M., Thorpe, A. N., and Alterescu, S.: Anisotropic Electromagnetic Features of a Grain-Aligned $YBa_2Cu_3O_x$ Bulk Superconductor. Phys. Rev. B, accepted for publication.

10. Swartzendruber, L. J., Roitburd, A., Kaiser, D. L., Gayle, F. W., and Bennett, L.: Direct Evidence for an Effect of Twin Boundaries on Flux Pinning in Single-Crystal $YBa_2Cu_3O_{6+x}$. Phys. Rev. Lett. Vol. 64, No. 4, 1990, pp. 483-486.

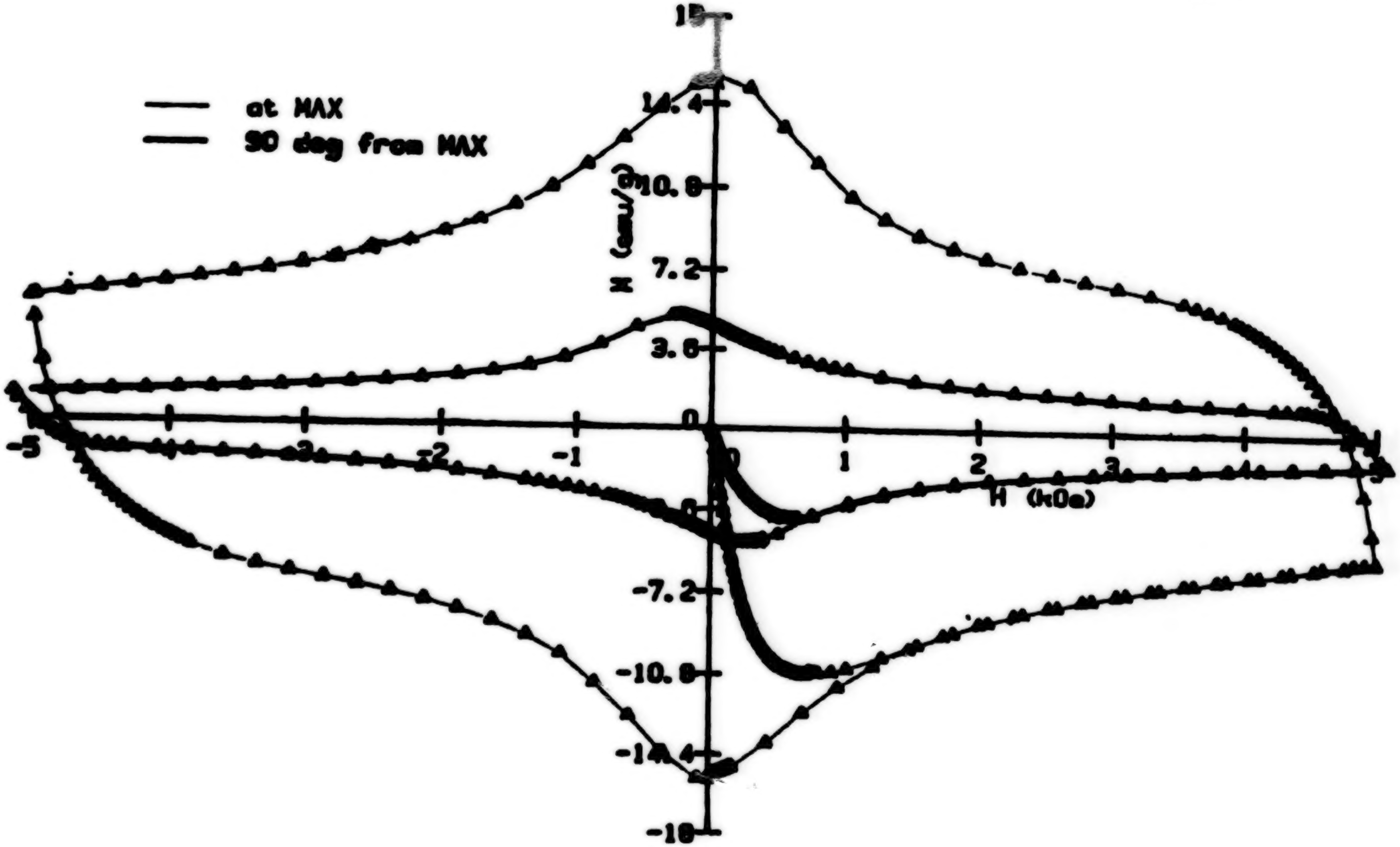
11. Savvides, N.: Flux Creep and Transport Critical Current Density in High-T_c Superconductors. Physica. C Vol. 165, Nos. 5-6, 1990, pp. 371-376.

12. Anderson, P. W.: Theory of Flux Creep in Hard Superconductors. Phys. Rev. Lett. Vol. 9, No. 7, 1962, pp. 309-311.

13. Hagen, C. W., Griessen, R. P., and Salomons, E.: Thermally Activated Flux Motion in High-T_c Superconductors: An Analytical Model. Physica. C Vol. 157, No. 2, 1989, pp. 199-208.

14. Mohamed, M. A.-K., Jung, J., and Franck, J. P.: Flux Creep and Penetration in Fe-doped $YBa_2Cu_3O_7$. Phys. Rev. B Vol. 41, No. 7, 1990, pp. 4286-4295.

Figure 1



T=LN2 mag moment=10; delay=100sec

Figure 2

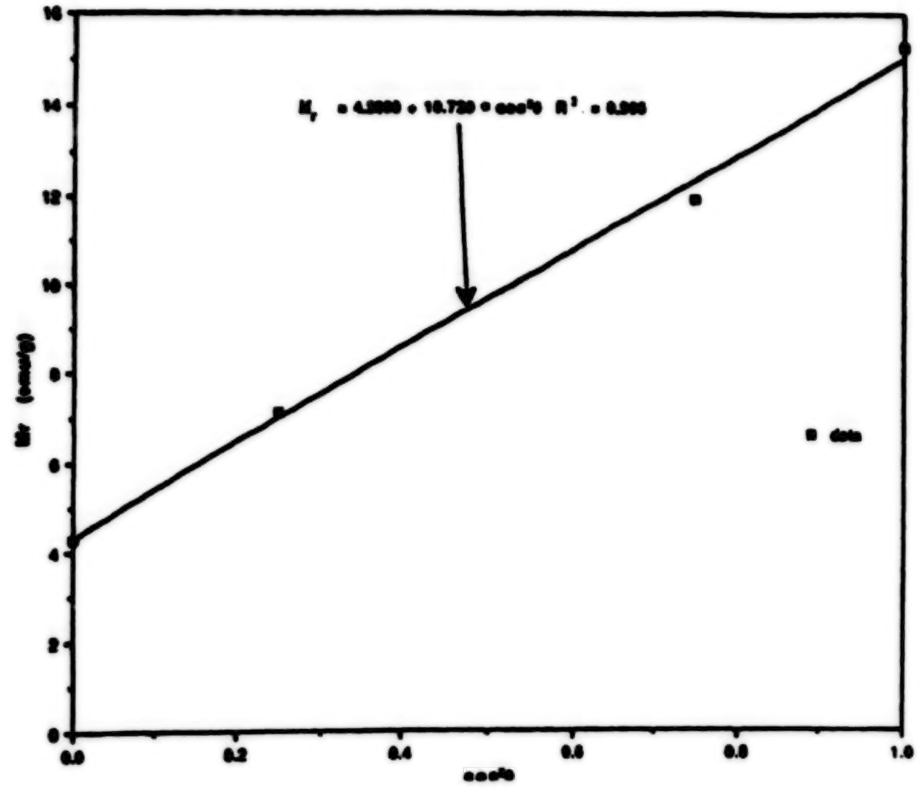


Figure 3

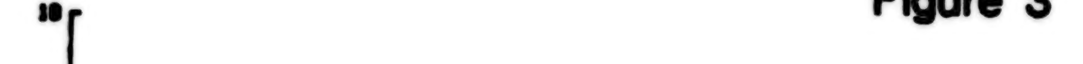


Figure 4

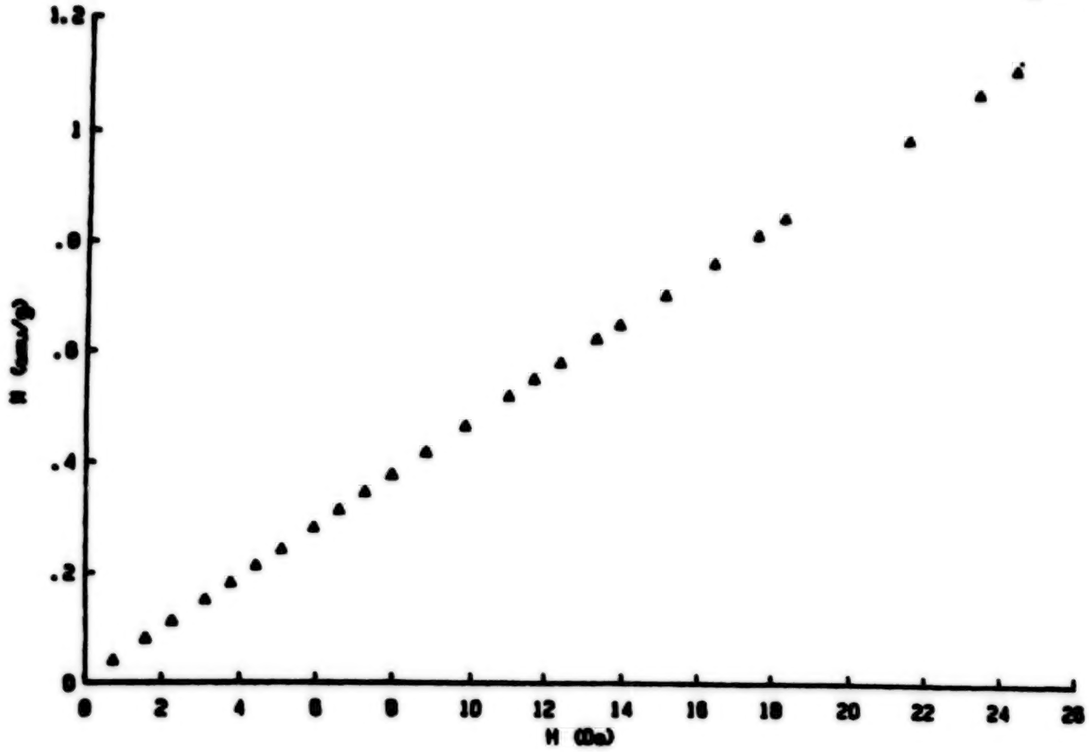
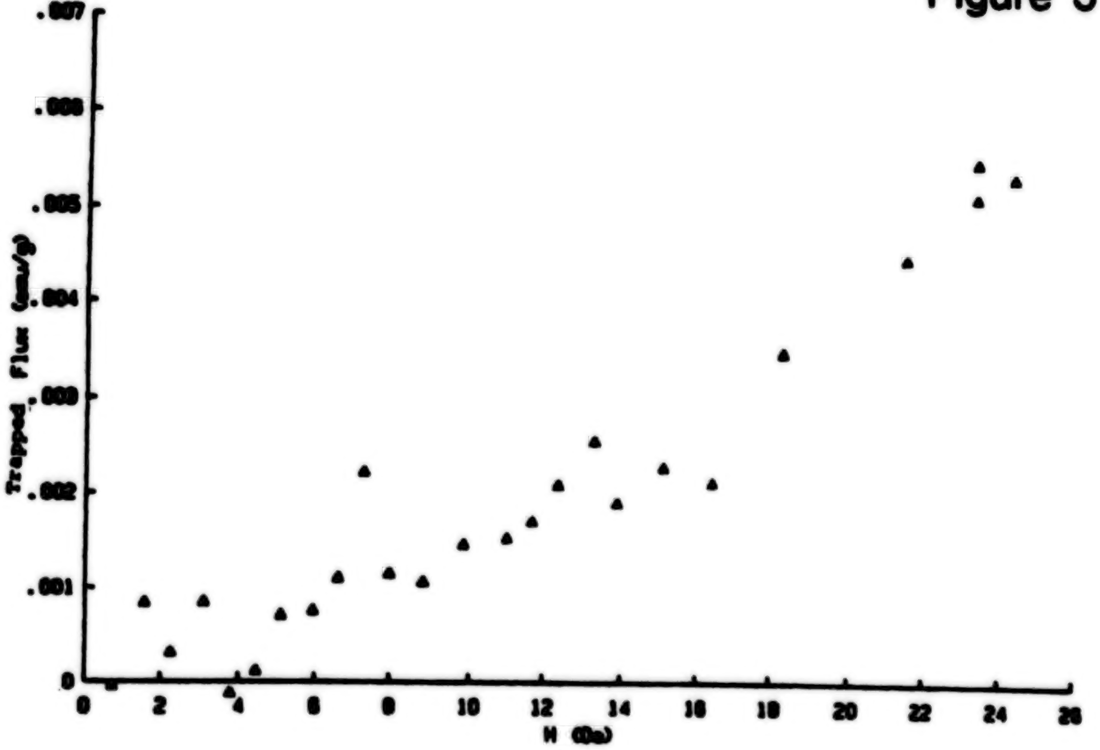


Figure 5



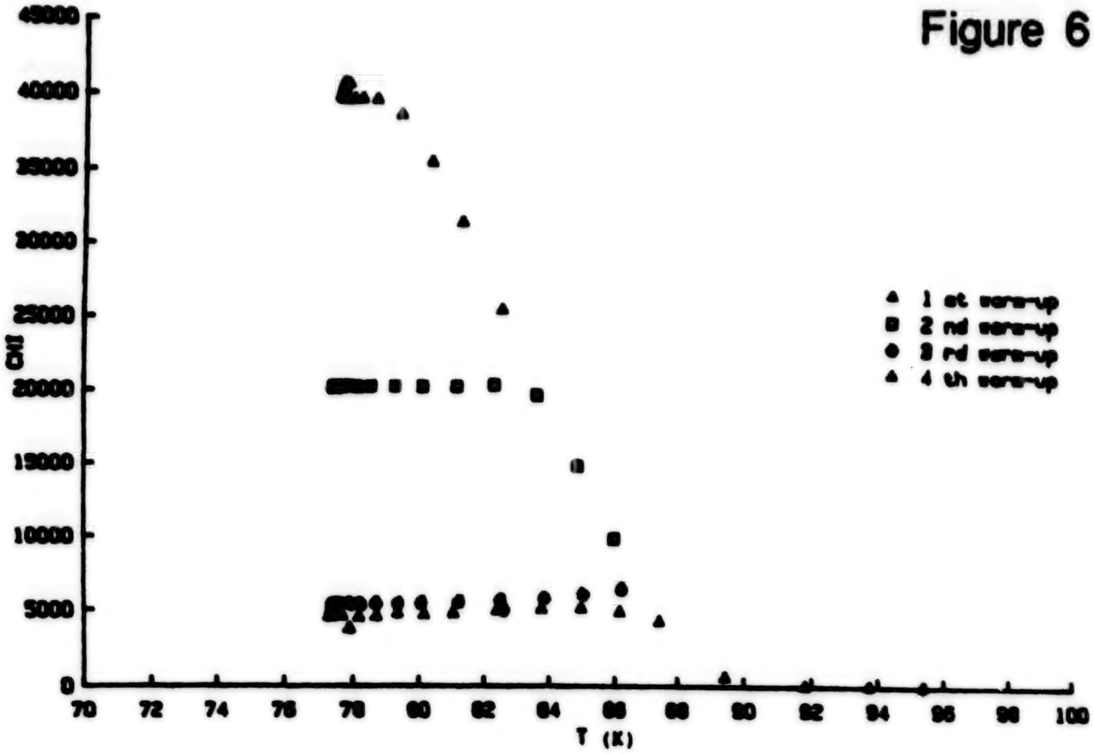


Figure 6

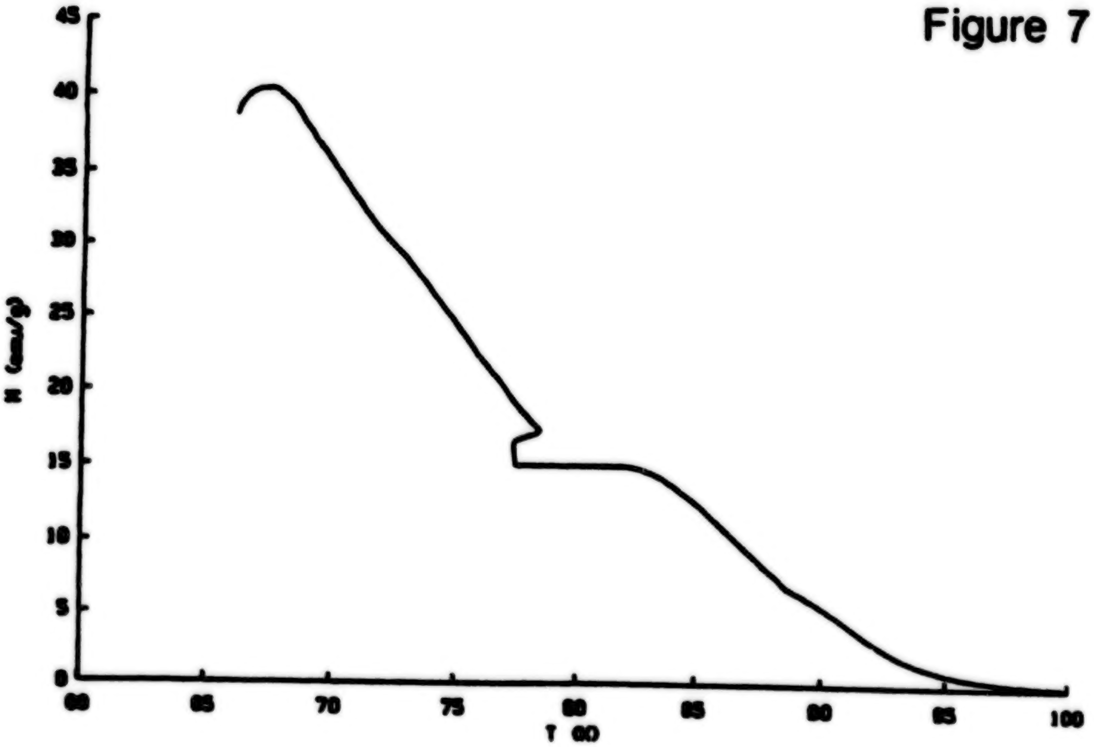
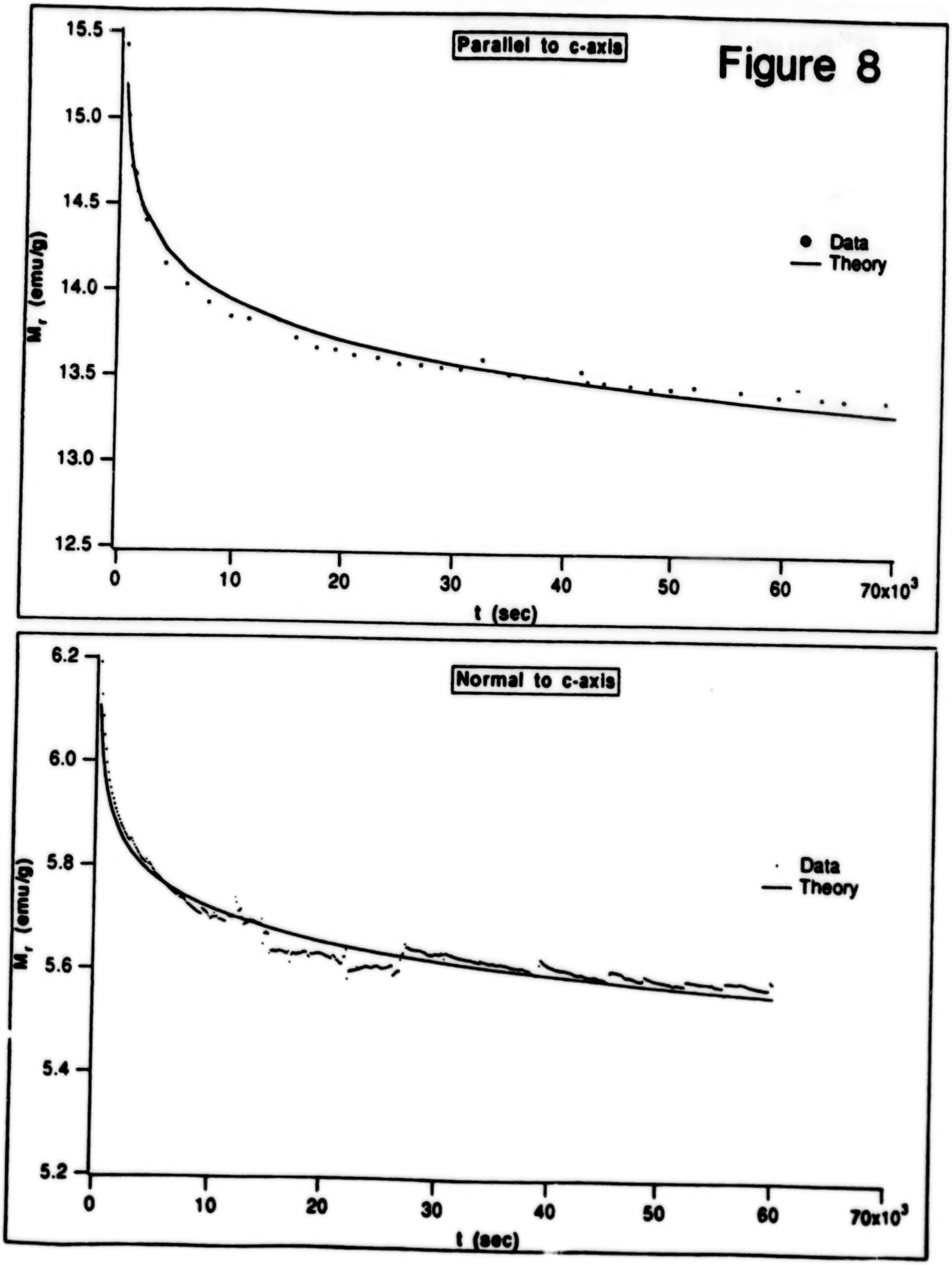


Figure 7

Figure 8


Parallel to c-axis
M_r (emu/g)
15.5
15.0
14.5
14.0
13.5
13.0
12.5
0
10
20
30
40
50
60
70x10^3
t (sec)
● Data
— Theory
Normal to c-axis
M_r (emu/g)
6.2
6.0
5.8
5.6
5.4
5.2
0
10
20
30
40
50
60
70x10^3
t (sec)
· Data
— Theory

N91
24083
UNCLAS

N91-24083

DESCRIPTION OF THE LARGE GAP MAGNETIC SUSPENSION SYSTEM (LGMSS) GROUND-BASED EXPERIMENT

By

Nelson J. Groom

Mail Stop 161
NASA Langley Research Center
Hampton, VA 23665

SUMMARY

A description of the Large Gap Magnetic Suspension System (LGMSS) ground-based experiment is presented. The LGMSS provides five-degrees-of-freedom control of a cylindrical suspended element which is levitated above a floor-mounted array of air core electromagnets. The uncontrolled degree of freedom is rotation about the long axis of the cylinder (roll). Levitation and control forces are produced on a permanent magnet core which is embedded in the cylinder. The cylinder also contains light emitting diodes (LEDs) and associated electronics and power supply. The LEDs provide active targets for an optical position measurement system which is being developed in-house at Langley Research Center. The optical position measurement system will provide six-degrees-of-freedom position information for the LGMSS control system.

INTRODUCTION

This paper describes the Large Gap Magnetic Suspension System (LGMSS) ground-based experiment and presents a simplified analytical model which can be used in analyses and simulations in the development of control system approaches and in evaluations of overall systems performance. The objectives of the LGMSS ground-based experiment are to investigate the technology issues associated with magnetic suspension, accurate suspended element control, and accurate position sensing at large gaps. This technology has potential applications in a wide range of areas including microgravity and vibration isolation systems, magnetically suspended pointing mounts, large-angle magnetic suspension systems for advanced actuators, wind tunnel magnetic suspension systems, and remote manipulation/control/positioning of objects in space. The simplified analytical model is based on the model developed in reference 1. This model was used to investigate candidate control approaches for the LGMSS. The control approaches are described and numerical results presented in reference 2.

EXPERIMENT DESCRIPTION AND BACKGROUND

The LGMSS ground-based experiment, as originally defined, is shown schematically in figure 1. It consists of a cylindrical suspended element which has a core composed of permanent magnet material embedded in it. Levitation forces and control forces and torques are produced on the permanent magnet core by air core electromagnets which are required to fit within an eight foot by eight foot square by four foot high volume. The core is suspended a total distance of three feet above the top surface of the electromagnet volume. In addition to the permanent magnet core, the suspended element also contains an array of LEDs and associated electronics and power supply. The LEDs are embedded in the surface of the suspended element and provide active targets for a photogrammetric optical position measurement system which is being developed at Langley Research Center. Each LED target is imaged by a cylindrical lens on a linear Charge Coupled Device (CCD) sensor. Position and orientation of the model is determined from the position of the projected target images. There are two sensors per sensing unit and a total of eight sensing units which are positioned symmetrically about and approximately six feet above the suspended element. The optical position measurement system provides six-degrees-of-freedom position information for the control system. The original parameters of the experiment are given in table 1.

Feasibility Studies

After defining the LGMSS experiment, two studies were performed to verify the feasibility of building a system to meet the experiment requirements and to investigate approaches to implement it. One study was performed by Madison Magnetics, Inc. and resulted in a proposed configuration of five electromagnets mounted in a planar array (ref. 3). This approach was designated the five-coil system and the study results are summarized below. The other study was performed by SatCon Technology Corporation and resulted in a proposed configuration of six electromagnets mounted in a planar array. This approach was designated the six-coil system and the study results are also summarized below.

Five-coil system.- An important conclusion of the Madison Magnetics study was that the implementation of the LGMSS experiment was feasible. The proposed implementation is shown schematically in figure 2 and consists of a planar array of five electromagnets mounted in a circular configuration. Since the LGMSS requirement is for five-degrees-of-freedom control, this represents the minimum number of actuators. The electromagnets are conventional liquid-helium cooled superconductors and combine the functions of levitation and control. The magnetization vector is horizontal (parallel to the long axis of the core) and the system is capable of providing 360 degrees yaw (rotation about the vertical axis) control.

Six-coil system.- The SatCon study also concluded that it was feasible to implement the LGMSS experiment. Their proposed approach is shown in figure 3 and consists of a planar array of six electromagnets mounted in a circular configuration. The two approaches are similar with the major differences being in the control approach and the number of coils. The six-coil configuration also uses electromagnets which are conventional liquid-helium cooled superconductors and which combine the functions of levitation and control. The magnetization vector is horizontal (parallel to the long axis of the core) and the system is capable of providing 360 degrees yaw control. The main reasons for adding a sixth coil were control system related. The six coil system results in a symmetrical configuration and also results in an overspecified system from the standpoint of control inputs. The sixth coil could be fitted in the allowable volume without a significant increase in total Ampere-turns.

Selected Configuration

As a result of the feasibility studies and further in-house studies, the requirements for the LGMSS experiment were refined and a decision was made to procure the design, fabrication, installation, and test of an LGMSS. The revised LGMSS requirements are presented in table 2. An open, competitive, procurement effort resulted in the selection of a configuration proposed by Intermagnetics General Corporation. This configuration is shown schematically in figure 4. As shown in the figure, there are two large concentric levitation coils and a separate set of control coils. The levitation coils are superconducting coils which are operated in the persistent mode. In the persistent mode, a superconducting coil is charged up to a certain current value and the terminals are shorted through a persistent mode switch. Since the superconductor has zero resistance, the current continues to flow or persist in the coil. In the configuration shown, the coils have currents flowing in opposite directions. The control coils are shown in a generic configuration since the contract is in the design phase and a final configuration has not been selected. The control coils are conventional room temperature coils. Figure 5 shows the levitation coils and permanent magnet core in more detail. As shown in the figure, the magnetization vector of the core is vertical (perpendicular to the long axis). By adjusting the persistent-mode currents to the correct values, a vertical field and gradient can be produced at the location of the core which will produce a stable levitation force and also a stable torque about the roll and pitch axes. Required control forces and torques are provided by the separate control coils. Yaw torque in this configuration is provided by producing a second-order gradient (gradient of a gradient) along the long axis of the core in the x-y plane. It should be noted that this configuration has the potential for providing active roll control.

LGMSS ANALYTICAL MODEL

This section presents an analytical model of the LGMSS which is based on the model developed in reference 1. This model assumes a magnetization vector which is parallel to the long axis of the core. An analytical model for a core with a vertical magnetization vector, which includes yaw torque generation with second-order gradients, is being developed. Figure 1 shows the coordinate systems and initial alignment. A set of orthogonal $\bar{x}, \bar{y}, \bar{z}$ body fixed axes defines the motion of the core with respect to inertial space. The core coordinate system is initially aligned with an orthogonal x, y, z system fixed in inertial space. A set of orthogonal x_b, y_b, z_b axes, also fixed in inertial space, define the location of the electromagnet array with respect to the x, y, z system. The x_b and y_b axes are parallel to the x and y axes respectively and the z_b and z axes are aligned. The centers of the two axis systems are separated by the distance h. The angular acceleration of the core, in core coordinates, can be written as (see ref. 1)

$$\{\dot{\bar{\Omega}}\} = (1/I_c)(\mathrm{Vol}(\{\bar{M}\} \text{ X } [T_m]\{B\}) + \{\bar{T}_d\}) \tag{1}$$

where I_c is the core moment of inertia about the axes of symmetry (y and z), Vol is the volume of the core, $\{\bar{M}\}$ is the magnetization of the core, $[T_m]$ is the vector transformation matrix from inertial to core coordinates, $\{B\}$ is the flux density produced by the electromagnets, and $\{\bar{T}_d\}$ represents external disturbance torques. A bar over a variable indicates that it is referenced to core coordinates. The translational acceleration of the core, in core coordinates, can be written as

$$\{\dot{\bar{V}}\} = (1/m_c)(\mathrm{Vol}([T_m][\partial B][T_m]^{-1}\{\bar{M}\}) + \{\bar{F}_d\}) \tag{2}$$

where m_c is the mass of the core, $[\partial B]$ is a matrix of the gradients of $\{B\}$, and $\{F_d\}$ represents external disturbance forces in core coordinates. The simplified field model of reference 1 has been extended to include expansion of the field around the operating point to first order terms. This means that each element of $\{B\}$, for example B_x, can be written in the form

$$B_x = \tilde{B}_x + (\tilde{B}_{xx})x + (\tilde{B}_{xy})y + (\tilde{B}_{xz})z \tag{3}$$

where $\tilde{B}_x$, $\tilde{B}_{xx}$, $\tilde{B}_{xy}$, and $\tilde{B}_{xz}$ are values calculated at the operating point and the notation $\partial B_i/\partial j = B_{ij}$ has been used. Since the fields and gradients are linear functions of coil currents, the components of $\tilde{B}_x$ produced by coil n of an n-coil system can be written as

$$\tilde{B}_{xn} = K_{xn}(I_n/I_{max}) \tag{4}$$

where I_{max} is the maximum coil current, K_{xn} is a constant which represents the magnitude of $\tilde{B}_{xn}$ produced by I_{max}, and I_n is the coil current. For the total system, $\tilde{B}_x$ can be written as (see ref. 1)

$$\tilde{B}_x = (1/I_{max})\lfloor K_x \rfloor \{I\} \tag{5}$$

where

$$\lfloor K_x \rfloor = \lfloor K_{x1}\ K_{x2} \bullet\bullet\bullet\ K_{xn} \rfloor \tag{6}$$

and

$$\{I\}^T = \lfloor I_1\ I_2 \bullet\bullet\bullet\ I_n \rfloor \tag{7}$$

The other fields and gradients can be written in the same way. A block diagram of the system is shown in figure 6. This model is nonlinear and is of the form

$$\dot{x} = f(x,u) \tag{8}$$

where x is given by

$$x^T = \lfloor \bar{\Omega}_y\ \ \bar{\Omega}_z\ \ \Theta_y\ \ \Theta_z\ \ \bar{v}_x\ \ \bar{v}_y\ \ \bar{v}_z\ \ x\ \ y\ \ z \rfloor \tag{9}$$

and the input u is given by

$$u^T = \lfloor I_1\ I_2 \bullet\bullet\bullet\ I_n \rfloor \tag{10}$$

The states Θ_y and Θ_z in (9) are the pitch and yaw angles of the core respectively.

CONCLUDING REMARKS

The Large Gap Magnetic Suspension System (LGMSS) ground-based experiment has been described and a simplified analytical model presented. The analytical model is for a suspended element core with a horizontal magnetization vector. An analytical model for a suspended element core with a vertical magnetization vector, which includes yaw torque generation with second-order gradients, is being developed. The objectives of the experiment are to investigate the technology issues associated with magnetic suspension, accurate suspended element control, and accurate position sensing at large gaps. This technology has potential applications in a wide range of areas including microgravity and vibration isolation systems, magnetically suspended pointing mounts, large-angle magnetic suspension systems for advanced actuators, wind tunnel magnetic suspension systems, and remote manipulation/control/positioning of objects in space.

REFERENCES

1. Groom, Nelson J.: Analytical Model Of A Five Degree Of Freedom Magnetic Suspension And Positioning System. NASA TM-100671, March 1989.

2. Groom, Nelson J.; and Schaffner, Philip R.: An LQR Controller Design Approach for a Large Gap Magnetic Suspension System (LGMSS). NASA TM-101606, July 1990.

3. Boom, R.W.; Abdelsalam, M.K.; Eyssa, Y.M.; and McIntosh, G.E.: Repulsive Force Support System Feasibility Study. NASA CR-178400, October, 1987.

TABLE 1. - LGMSS FEASIBILITY STUDY PARAMETERS

- Core dimensions
  (two sizes)

  - Diameter = 4 in.
    Length = 9 in.

  - Diameter = 2 in.
    Length = 12 in.

- Suspension support capability
  - Core weight plus 15 lb.

- Electromagnet volume
  (floor mounted)

  - 8 ft. X 8 ft. square X 4 ft. high

- Suspension height
  (above electromagnet volume)

  - 3 ft.

- Position range
  (yaw)

  - 40 deg.

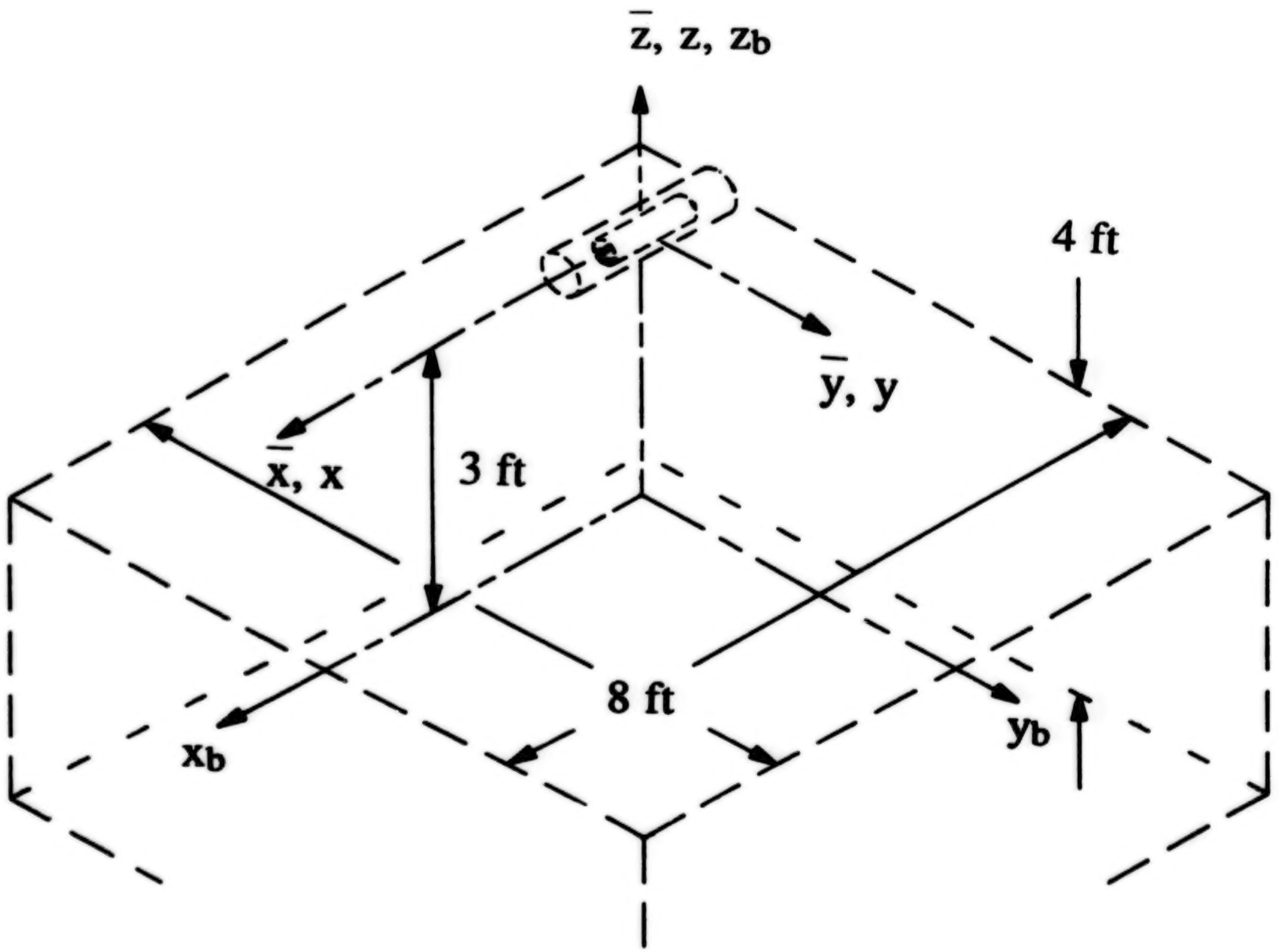
- Accuracy

  - Translation (x, y, z) = ± 0.001 in.

  - Rotation ($\Theta_x, \Theta_y, \Theta_z$) = ± 0.002 deg.

TABLE 2. - LGMSS STATEMENT OF WORK PARAMETERS

- Core dimensions
  (two sizes)
  - Diameter = 1.25 in.
    Length = 5.85 in.
  - Diameter = 2.30 in.
    Length = 7.80 in.
- Suspension support capability
  - Core weight plus 50% of core weight
- Electromagnet volume
  (floor mounted)
  - 8 ft. X 8 ft. square X 4 ft. high
- Suspension height
  (above electromagnet volume)
  - 3 ft.
- Position range
  (yaw)
  - ±360 deg.
- Accuracy
  - Translation (x, y, z) = ± 0.01 in.
  - Rotation (Θ_x, Θ_y, Θ_z) = ± 0.02 deg.



$\bar{z}$, z, z_b
4 ft
$\bar{y}$, y
$\bar{x}$, x
3 ft
8 ft
x_b
y_b

Figure 1.- LGMSS ground-based experiment definition.

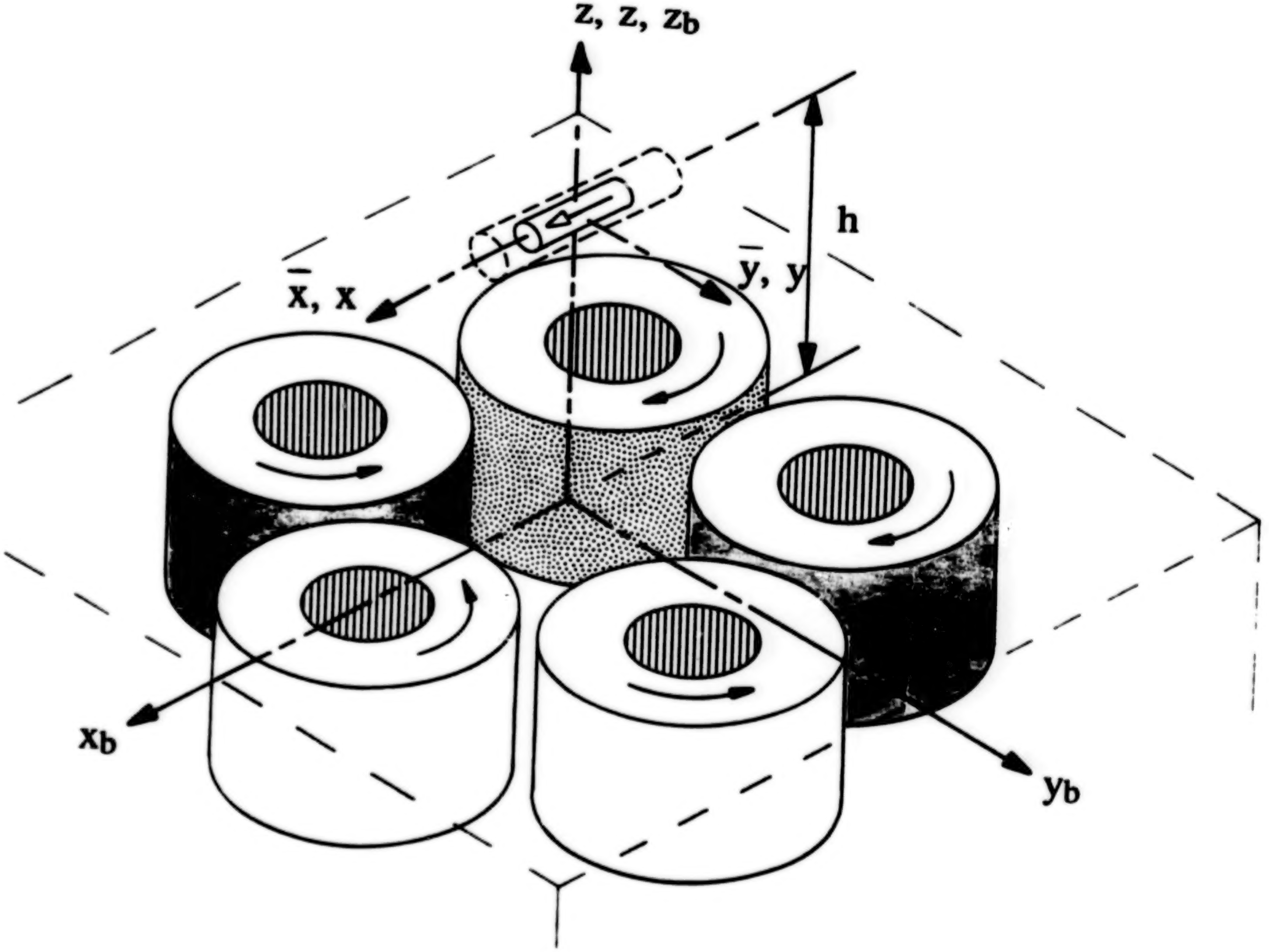
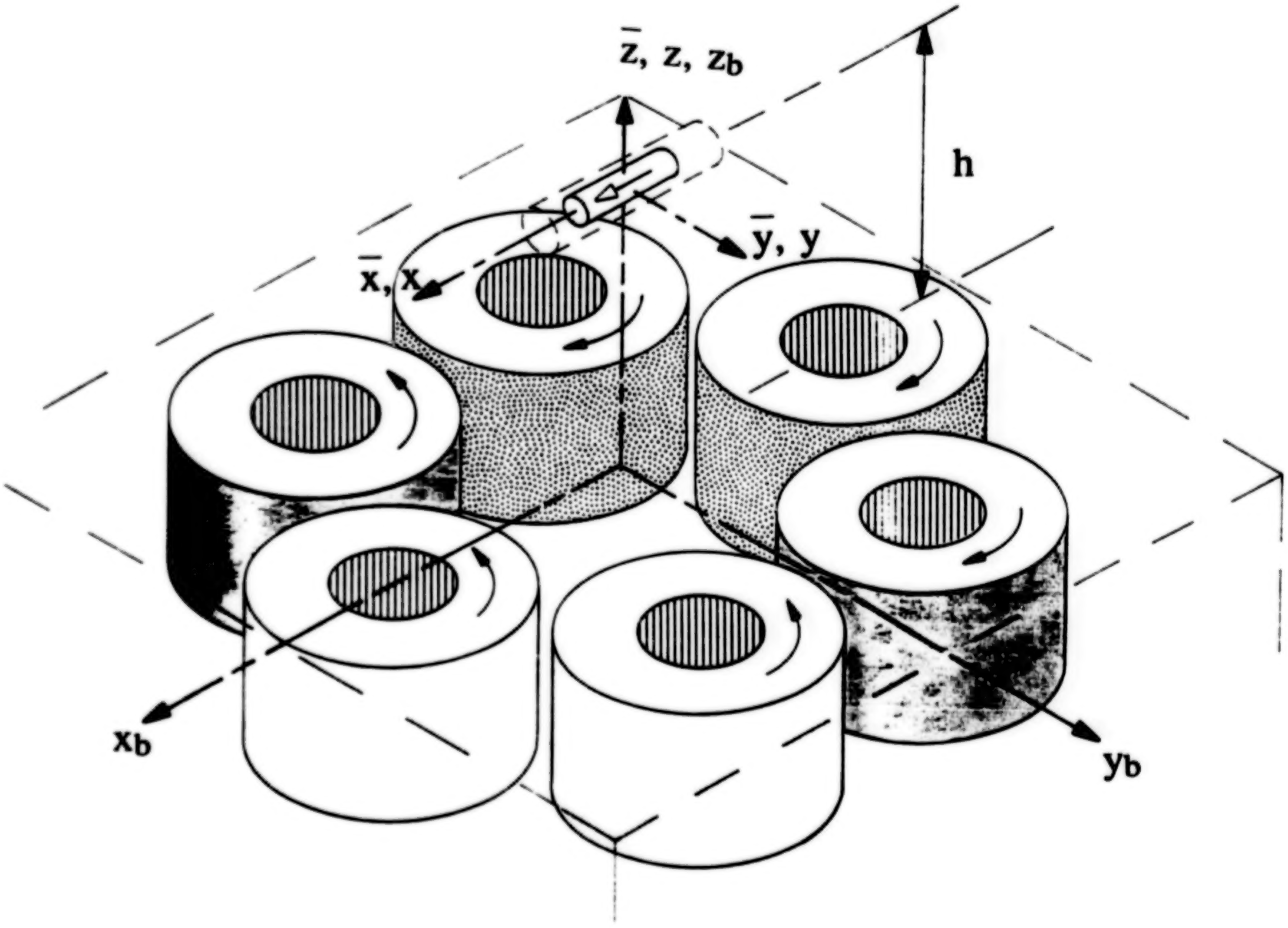

$\bar{z}, z, z_b$
h
$\bar{x}, x$
$\bar{y}, y$
x_b
y_b

Figure 2.- LGMSS Five-Coil configuration.



$\bar{z}$, z, z_b
h
$\bar{y}$, y
$\bar{x}$, x
x_b
y_b

Figure 3.- LGMSS Six-Coil configuration.

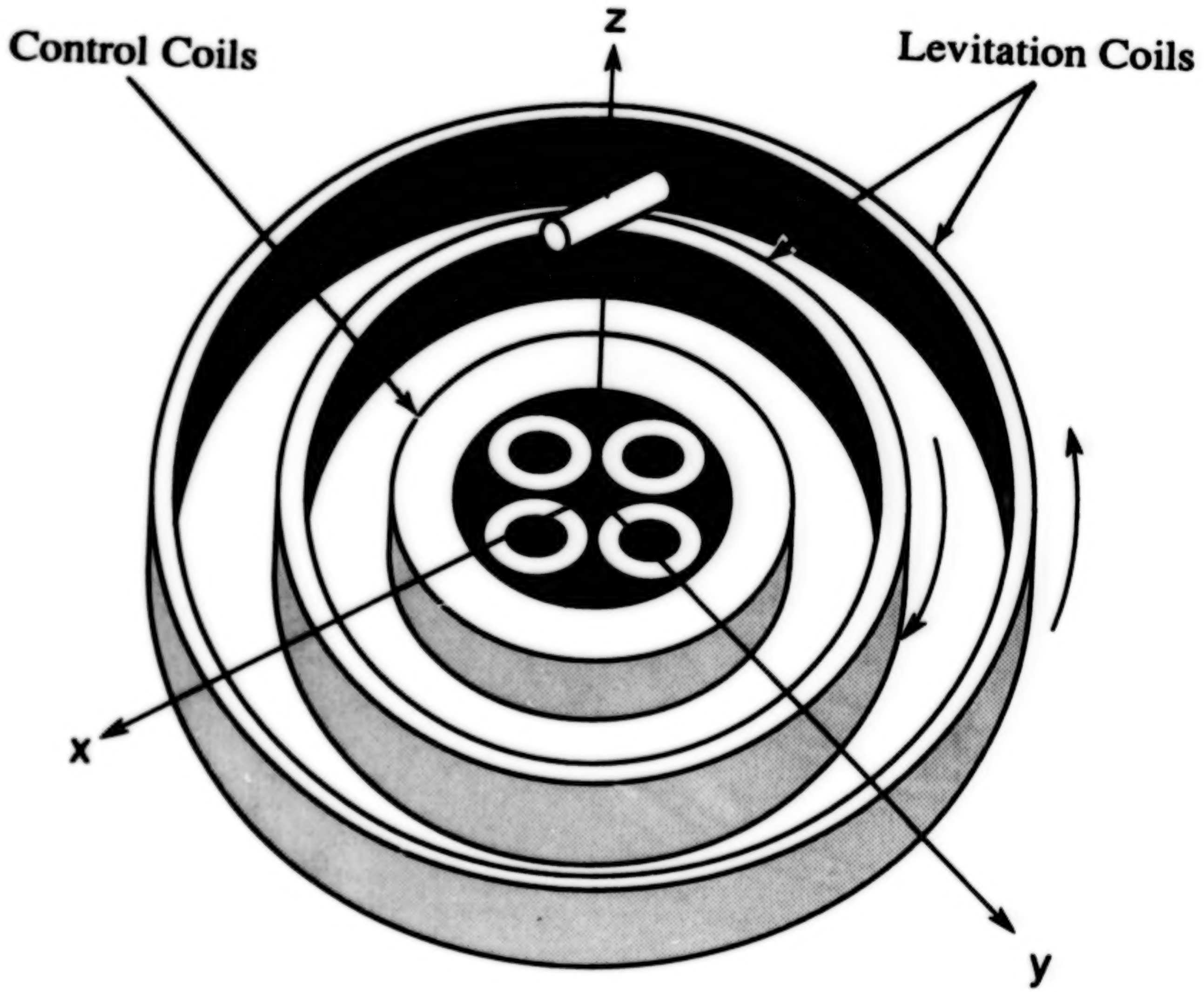

Control Coils
z
Levitation Coils
x
y

Figure 4.- Selected configuration for LGMSS ground-based experiment.

109.1213.001.107.A

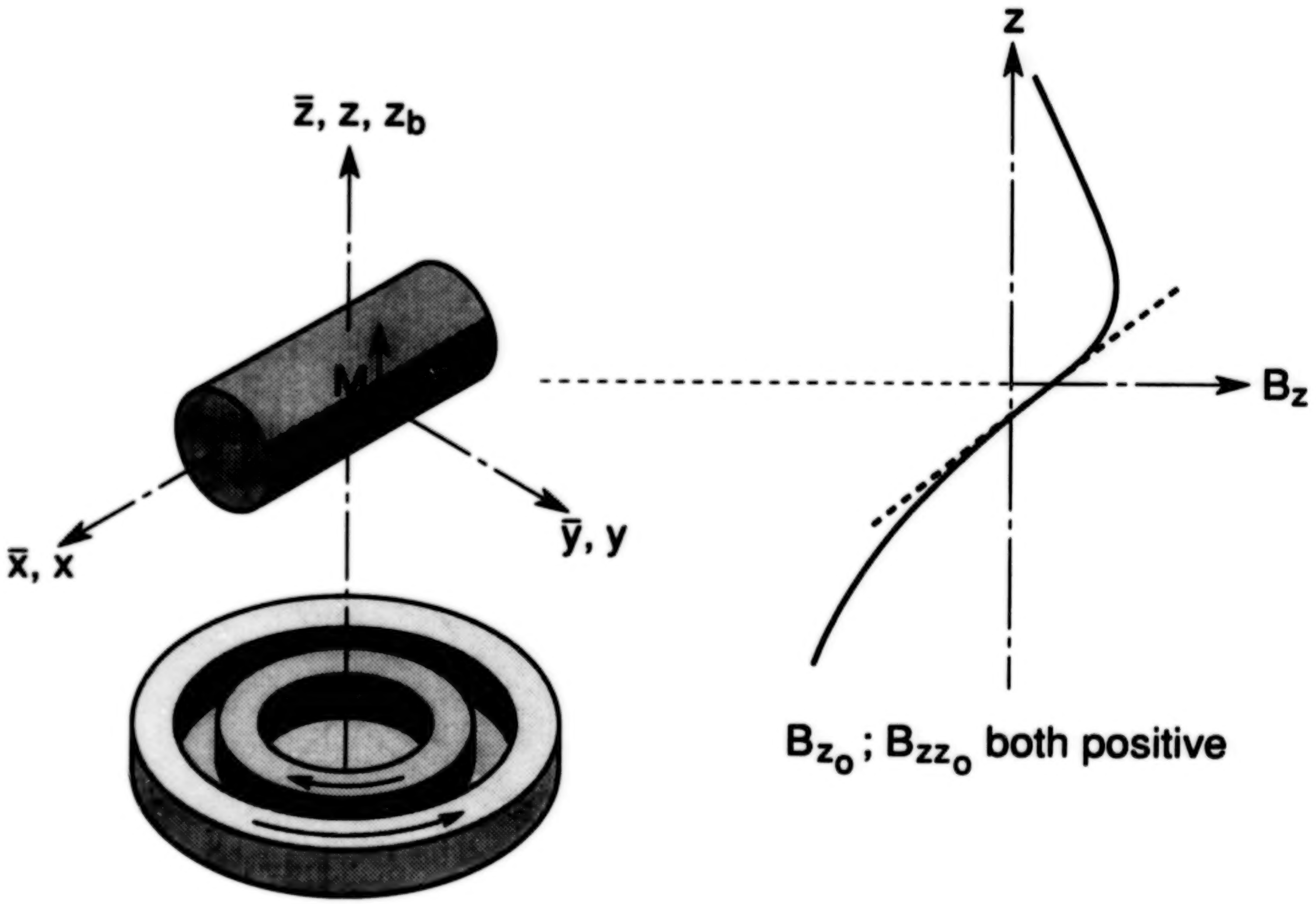

$\bar{z}$, z, z_b
M
$\bar{x}$, x
$\bar{y}$, y
z
B_z
B_{z_o} ; B_{zz_o} both positive

Figure 5.- Permanent magnet core and levitation coil detail for selected LGMSS configuration.

109.1213.001.108.A

☆ U.S. GOVERNMENT PRINTING OFFICE: 1991-527-064/60016

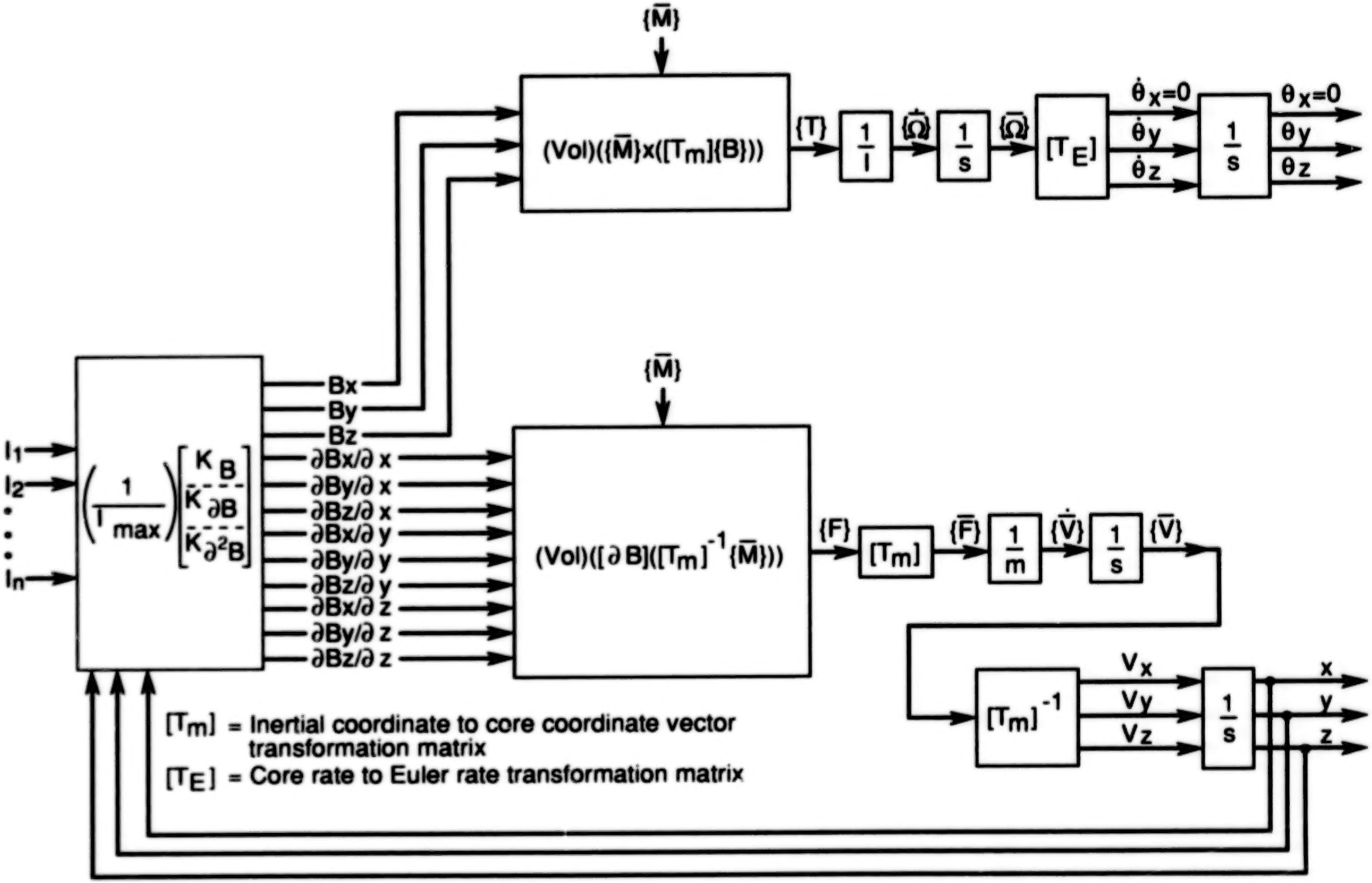

{M̄}
(Vol)({M̄}x([Tm]{B}))
{T}
1/I
{Ω̇}
1/s
{Ω}
[TE]
θ̇x=0
θ̇y
θ̇z
1/s
θx=0
θy
θz
I1
I2
In
(1/Imax)[KB / K∂B / K∂²B]
Bx
By
Bz
∂Bx/∂x
∂By/∂x
∂Bz/∂x
∂Bx/∂y
∂By/∂y
∂Bz/∂y
∂Bx/∂z
∂By/∂z
∂Bz/∂z
{M̄}
(Vol)([∂B]([Tm]^-1{M̄}))
{F}
[Tm]
{F̄}
1/m
{V̇}
1/s
{V̄}
[Tm]^-1
Vx
Vy
Vz
1/s
x
y
z
[Tm] = Inertial coordinate to core coordinate vector transformation matrix
[TE] = Core rate to Euler rate transformation matrix

Figure 6.- Block diagram of LGMSS analytical model (horizontal magnetization vector).

109.1213.001.109 A

National Aeronautics and Space Administration

Report Documentation Page

1. Report No. NASA CP-3109, Vol. 2

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle Technology 2000

5. Report Date March 1991

6. Performing Organization Code CU

7. Author(s)

8. Performing Organization Report No.

9. Performing Organization Name and Address
NASA Technology Utilization Division

10. Work Unit No.

11. Contract or Grant No.

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Washington, DC 20546

13. Type of Report and Period Covered Conference Publication

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

Proceedings from the Technical Sessions of the Technology 2000 Conference and Exposition, November 27-28, 1990, Washington, DC. Volume 2 features 46 papers presented during eight concurrent sessions.

17. Key Words (Suggested by Author(s))
technology transfer
artificial intelligence
environmental technology
superconductivity

18. Distribution Statement
Unclassified - Unlimited

Subject Category 99

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	388	A17

NASA FORM 1626 OCT 86

END

DATE

FILMED

JUL 5 1991